Session 9: Introduction to Sieve Analysis of Pathogen Sequences, for Assessing How VE Depends on Pathogen Genomics– Part I

Peter B Gilbert

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington

June 22, 2016
### Outline of Module 8: Evaluating Vaccine Efficacy

<table>
<thead>
<tr>
<th>Session</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Halloran)</td>
<td>Introduction to Study Designs for Evaluating VE</td>
</tr>
<tr>
<td>2 (Follmann)</td>
<td>Introduction to Vaccinology Assays and Immune Response</td>
</tr>
<tr>
<td>3 (Gilbert)</td>
<td>Introduction to Frameworks for Assessing Surrogate Endpoints/Immunological Correlates of VE</td>
</tr>
<tr>
<td>4 (Follmann)</td>
<td>Additional Study Designs for Evaluating VE</td>
</tr>
<tr>
<td>5 (Gilbert)</td>
<td>Methods for Assessing Immunological Correlates of Risk and Optimal Surrogate Endpoints</td>
</tr>
<tr>
<td>6 (Gilbert)</td>
<td>Effect Modifier Methods for Assessing Immunological Correlates of VE (Part I)</td>
</tr>
<tr>
<td>7 (Gabriel)</td>
<td>Effect Modifier Methods for Assessing Immunological Correlates of VE (Part II)</td>
</tr>
<tr>
<td>8 (Sachs)</td>
<td>Tutorial for the R Package <code>pseval</code> for Effect Modifier Methods for Assessing Immunological Correlates of VE</td>
</tr>
<tr>
<td><strong>Session 9 (Gilbert)</strong></td>
<td><strong>Introduction to Sieve Analysis of Pathogen Sequences, for Assessing How VE Depends on Pathogen Genomics</strong></td>
</tr>
<tr>
<td>10 (Follmann)</td>
<td>Methods for VE and Sieve Analysis Accounting for Multiple Founders</td>
</tr>
</tbody>
</table>
Figure 1 from Gilbert, Self, Ashby (1998, Biometrics)
Outline of Session 9

1. Sieve Analysis Via Cumulative and Instantaneous VE Parameters
2. Cumulative VE Approach: NPMLE and TMLE
3. Mark-Specific Proportional Hazards Model
4. Example 1: RV144 HIV-1 Vaccine Efficacy Trial
5. Example 2: RTS,S Malaria Vaccine Efficacy Trial
Cumulative Genotype-Specific VE

- $T = \text{time from study entry (or post immunization series) until study endpoint through to time } \tau_1$ (e.g., HIV-1 infection)
- $t = \text{fixed time point of interest } t < \tau_1$

- **Discrete** genotype-specific cumulative VE

$$VE_{cml/disc}^{\text{cm}}(t, j) = \left[ 1 - \frac{P(T \leq t, J = j|\text{Vaccine})}{P(T \leq t, J = j|\text{Placebo})} \right] \times 100\%,$$ $t \in [0, \tau_1]$

- **Continuous** genetic distance-specific cumulative VE

$$VE_{cml/cont}^{\text{cm}}(t, v) = \left[ 1 - \frac{P(T \leq t, V = v|\text{Vaccine})}{P(T \leq t, V = v|\text{Placebo})} \right] \times 100\%,$$ $t \in [0, \tau_1]$

- $J = \text{discrete genotype subgroup such as binary, unordered categorical, ordered categorical}$
- $V = \text{(approximately) continuous genetic distance to a vaccine sequence}$
Cumulative $\textit{VE}$ Sieve Effect Tests

Fix $t$ at the primary time point of interest

- $\text{VE}^{cml/disc}(t,j)$:

  $H_0 : \text{VE}^{cml/disc}(t,j)$ constant in $j$
  $H_1^{\text{mon}} : \text{VE}^{cml/disc}(t,j)$ decreases in $j$
  $H_1^{\text{any}} : \text{VE}^{cml/disc}(t,j)$ has some differences in $j$

- $\text{VE}^{cml/cont}(t,v)$:

  $H_0 : \text{VE}^{cml/cont}(t,v)$ constant in $v$
  $H_1^{\text{mon}} : \text{VE}^{cml/cont}(t,v)$ decreases in $v$
  $H_1^{\text{any}} : \text{VE}^{cml/cont}(t,v)$ has some differences in $v$

A “sieve effect” is defined by $H_1^{\text{mon}}$ or $H_1^{\text{any}}$ being true (i.e., differential $\text{VE}$ by pathogen genotype)
Illustration: Cumulative $\text{VE}_{cml/disc}(t = 14, j)$ for 3-Level $J^*$

Discrete Genotype–Specific Cumulative VE at $t = 14$ Months

Illustration: Cumulative $VE^{cml/cont}(t = 14, \nu)$ for Continuous Distance $V^*$

Estimation of Cumulative VE Parameters: Approach Without Covariates

- Nonparametric maximum likelihood estimation and testing

Assumptions Required for Consistent Inference

- **No interference**: Whether a subject experiences the malaria endpoint does not depend on the treatment assignments of other subjects
- **A randomized trial**
- **Random dropout**: Whether a subject drops out by time $t$ does not depend on observed or unobserved subject characteristics
- **MCAR genotypes**: Endpoint cases with missing pathogen genomes have missingness mechanism Missing Completely at Random (MCAR)
Estimation of Cumulative VE Parameters: With Covariates

- Targeted minimum loss-based estimation (tMLE) and testing

Assumptions Required for Consistent Inference

- No interference
- A randomized trial
- Correct modeling of dropout
- Missing at Random genotypes

Advantages of approach with covariates

- Correct for bias due to covariate-dependent dropout
- Increase precision via covariates predicting the endpoint and/or dropout
- Correct for bias from covariate-dependent missing genotypes (e.g., pathogen load-dependent)
- Increase precision by predicting missing genotypes (the best predictors would be based on pathogen sequences of later-sampled pathogens)
• $h(t, j) = \text{Hazard of the malaria endpoint with discrete genotype } j$

• $\lambda(t, v) = \text{Hazard of the malaria endpoint with continuous genetic distance } v$

• **Discrete** genotype-specific instantaneous vaccine efficacy

\[
VE^{\text{haz/disc}}(t, j) = \left[1 - \frac{h(t, j|\text{Vaccine})}{h(t, j|\text{Placebo})}\right] \times 100%
\]

• **Continuous** genetic distance-specific instantaneous vaccine efficacy

\[
VE^{\text{haz/cont}}(t, v) = \left[1 - \frac{\lambda(t, v|\text{Vaccine})}{\lambda(t, v|\text{Placebo})}\right] \times 100%
\]

• **Proportional hazards assumption:** $VE^{\text{haz/disc}}(t, j) = VE^{\text{haz/disc}}(j)$ and $VE^{\text{haz/cont}}(t, v) = VE^{\text{haz/cont}}(v)$ for all $t \in [0, \tau_1]$
Illustration: Instantaneous $VE^{haz/disc}(j)$ for 3-Level $J^*$

![Graph of Discrete Genotype-Specific Instantaneous VE to 14 Months]

- No. Cases (V:P): 12:25
  - Full Match, Unadjusted: $0.76$, $0.52$ ($p=0.036$)
  - Near, Unadjusted: $0.54$, $0.12$ ($p=0.031$)
  - Distant, Unadjusted: $0.05$, $0.05$ ($p=0.036$)

  - Full Match, Unadjusted: $0.73$, $0.71$ ($p=0.03$)
  - Near, Unadjusted: $0.44$, $0.42$ ($p=0.11$)
  - Distant, Unadjusted: $0.45$, $0.41$ ($p=0.03$)

- No. Cases (V:P): 19:18
  - Full Match, Unadjusted: $0.69$, $0.42$ ($p=0.10$)
  - Near, Unadjusted: $0.45$, $0.42$ ($p=0.11$)
  - Distant, Unadjusted: $0.41$, $0.04$ ($p=0.023$)

Illustration: Instantaneous $VE^{ haz/cont}(v)$ for Continuous Distance $V^*$

*Juraska and Gilbert (2013, *Biometrics*): overall endpoint Cox model + semiparametric biased sampling model
Discussion of Instantaneous vs. Cumulative VE Approaches

• **Disadvantages:**
  - The instantaneous approach requires the extra assumption of proportional hazards (typically fails because of waning VE)
  - The VE parameters are hard to interpret under violation of proportional hazards
  - With currently available methods, cannot adjust for covariates without changing the target parameter to one that is not of main interest
    - Must rely on a random dropout assumption (cannot allow dropout to depend on covariates)
    - Cannot increase statistical power and precision by leveraging covariates, nor flexibly correct for accidental confounding

• **Advantages:**
  - If proportional hazards holds, the VE parameter is interpretable in terms of leaky genotype-specific vaccine efficacy
  - If proportional hazards approximately holds, may be reasonably interpretable and have increased efficiency by aggregating the vaccine efficacy over all time points
Outline of Session 9

1. Sieve Analysis Via Cumulative and Instantaneous VE Parameters
2. **Cumulative VE Approach: NPMLE and TMLE**
3. Mark-Specific Proportional Hazards Model
4. Example 1: RV144 HIV-1 Vaccine Efficacy Trial
5. Example 2: RTS,S Malaria Vaccine Efficacy Trial
Replace augmented IPW with TMLE (Benkeser, Carone, and Gilbert, 2016)
  - Unbiased under weaker assumptions; more efficient

The missing data methods assume a validation set—a subgroup of cases
where the founding pathogen genotype(s) is known with certainty
  - For pathogens that evolve very quickly post-infection (e.g., HIV-1), there may
    be no validation set!
  - Replace with measurement error methods, incorporating models predicting
    (imperfectly) founder HIV genotypes

Targeted learning approaches with **data adaptive genotype-specific VE**
target parameters that combine inference with model selection on the
marks/genotypes
Cumulative Genotype-Specific VE: Aalen-Johansen NPMLE

**Discrete genotype-specific cumulative VE**

\[ VE_{cml/disc}^{(t,j)} = \left[ 1 - \frac{P(T \leq t, J = j | \text{Vaccine})}{P(T \leq t, J = j | \text{Placebo})} \right] \times 100\%, \ t \in [0, \tau_1] \]

- Observe \( \tilde{T} \equiv \min(T, C) \) and \( \Delta J \equiv I(\tilde{T} = T)J \)
- With independent censoring, identify \( P(T \leq t, J = j | Z = z) \) via hazards:

\[
\bar{Q}^z_j(t) \equiv P(\tilde{T} = t, \Delta J = j | Z = z, \tilde{T} > t - 1)
\]

\[
\bar{Q}^z(t) \equiv \sum_{i=1}^{K} \bar{Q}^z_i(t)
\]

\[
P(T \leq t, J = j | Z = z) = \sum_{t'=1}^{t} \left[ \bar{Q}^z_j(t') \prod_{s=1}^{t'-1} \{1 - \bar{Q}^z(s)\} \right]
\]
Cumulative Genotype-Specific VE: Aalen-Johansen NPMLE

- Aalen-Johansen estimator plugs in empirical estimates

\[
\tilde{Q}_{j,n}^z(t) = \frac{\text{No. type j events at } t \text{ in group } z}{\text{No. at risk at } t-1 \text{ in group } z}
\]

\[
\hat{P}(T \leq t, J = j|Z = z) = \sum_{t' = 1}^{t} \left[ \tilde{Q}_{j,n}^z(t') \prod_{s=1}^{t'-1} \{1 - \tilde{Q}_{j,n}^z(s)\} \right]
\]

Limitations
- For consistency need random censoring (cannot depend on covariates)
- Efficient if no prognostic factors
Incorporating Covariates: TMLE

\[
P(T \leq t, J = j|Z = z) = E_W [P(T \leq t, J = j|Z = z, W)]
= \sum_{w} P(T \leq t, J = j|Z = z, W = w) P(W = w|Z = z)
\]

- TMLE optimizes bias-variance trade-off for estimating \( P(T \leq t, J = j|Z = z) \)
- Incorporates flexible models of \( P(T \leq t, J = j|Z = z, W) \) and of \( P(C \leq t|Z = z, W) \)
- TMLEs are doubly robust and asymptotically normal
  - Also asymptotically efficient if both \( P(T \leq t, J = j|Z = z, W) \) and \( P(C \leq t|Z = z, W) \) are estimated consistently
- Benkeser, Carone and Gilbert (2016) developed this TMLE, with R code
Mean Squared Error TMLE vs. Aalen-Johansen

Level of censoring

Covariate predictiveness of events

Covariate predictiveness of censoring

Relative mean squared error

0.7 0.9 1.1

Low

High

None

Med.

High

None

Med

High

None

0

Low

0

High

0.7 0.9 1.1

None Med. High None Med High

PBG (VIDD FHCRC) Sieve Analysis Methods June 22, 2016 20 / 37
Power of Wald Tests TMLE vs. Aalen-Johansen

Moderately prognostic covariates

Power, %

0 0.25 0.5 0.75

VE(t, j)

2.5 25 50 75 100

TMLE

Aalen−Johansen

Power relative to Aalen−Johansen

TMLE 1.00 1.02 1.03 1.01 1.00 1.00

20/28
Power of Wald Tests: TMLE vs. Aalen-Johansen

Strongly prognostic covariates

$VE(t, j)$

Power, %

0 0.25 0.5 0.75

2.5 25 50 75 100

TMLE

Aalen–Johansen

Power relative to Aalen–Johansen

TMLE

1.06 1.07 1.08 1.04 1.01 1.00
Sieve Analysis of RV144 Thai Trial

Background on Thai Trial

- Conducted 2004–2009 in the general population of Thailand
- 16,403 randomized 1:1 vaccine:placebo, primary endpoint HIV-1 infection by 3.5 years
- $\hat{VE} = 31\%$, 95% CI 1% to 51%, $p = 0.04$ (Rerks-Ngarm et al., 2009, *NEJM*)

\[ VE^{cml/disc}(3.5, \nu = 0) > VE^{cml/disc}(3.5, \nu = 1) \]

with \( V \) defined by match (\( \nu = 0 \)) vs. mismatch (\( \nu = 1 \)) of the infecting HIV-1 with the vaccine sequences at position 169 of HIV-1 Env V2

• TMLE adjusting for risky behaviors, gender, age, gave a similar result with increased precision (Benkeser, Carone, Gilbert, 2016); next slide
**AA position 169 matched**

\[ \overrightarrow{VE}_{\text{match}}(3.5) = 46\% \ (14\%, 66\%), \ p=0.01 \]

**AA position 169 mismatched**

\[ \overrightarrow{VE}_{\text{mismatch}}(3.5) = -39\% \ (-229\%, 42\%), \ p=0.46 \]
Outline of Session 9

1. Sieve Analysis Via Cumulative and Instantaneous VE Parameters
2. Cumulative VE Approach: NPMLE and TMLE
3. Mark-Specific Proportional Hazards Model
4. Example 1: RV144 HIV-1 Vaccine Efficacy Trial
5. Example 2: RTS,S Malaria Vaccine Efficacy Trial
Mark-Specific Proportional Hazards Approach with Missing Pathogen Sequences

- Gilbert and Sun (2015, *JRSS-B*)
- These methods pose a continuous mark-specific proportional hazards model and use inverse probability weighting (IPW) or augmented IPW
Competing Risks Model in Vaccine Efficacy Trials

- Conditional mark-specific hazard rate function:

\[ \lambda(t, v|z) = \lim_{h_1, h_2 \to 0} \frac{P\{T \in [t, t + h_1), V \in [v, v + h_2)| T \geq t, Z = z\}}{h_1 h_2} \]

- Covariate-adjusted mark-specific vaccine VE:

\[ \text{VE}(t, v|z) = 1 - \frac{\lambda_v(t, v|z)}{\lambda_p(t, v|z)}, \]

where \( \lambda_v(t, v|z) \) and \( \lambda_p(t, v|z) \) are the conditional mark-specific hazard functions for the vaccine and placebo groups, respectively
Mark-Specific Proportional Hazards Models

- Stratified mark-specific proportional hazards model:

\[
\lambda_k(t, v|z_{ki}(t)) = \lambda_{0k}(t, v)\exp\{\beta(v)^T z_{ki}(t)\}, \quad k = 1, \ldots, K
\]

where \(\lambda_{0k}(t, v)\) is an unspecified baseline function and \(\beta(v)\) is \(p\)-dimensional regression coefficient functions

- \(z = (z_1, z_2); \quad z_1 = \) vaccine group indicator; \(z_2\) other covariates; \(\beta_1(v) = \) coefficient corresponding to \(z_1\)

Mark-specific vaccine efficacy:

\[
VE(v) = 1 - \exp(\beta_1(v))
\]
Completely observed competing risks data:

\[(Z_{ki}, X_{ki}, \delta_{ki}, \delta_{ki}V_{ki}), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K,\]

where \(X_{ki} = \min\{T_{ki}, C_{ki}\}, \quad \delta_{ki} = I(T_{ki} \leq C_{ki})\)

When the failure time \(T_{ki}\) is observed, \(\delta_{ki} = 1\) and the mark \(V_{ki}\) is also observed, whereas if \(T_{ki}\) is censored, the mark \(V_{ki}\) is unknown.

Assume \(C_{ki}\) is independent of \(T_{ki}\) and \(V_{ki}\) conditional on \(Z_{ki}\).
Observed data

\[ O_{ki} = \{ X_{ki}, Z_{ki}, \delta_{ki}, R_{ki}, R_{ki} \delta_{ki} V_{ki}, \delta_{ki} A_{ki} \}, \quad i = 1 \ldots, n_k, \quad k = 1, \ldots, K, \]

\( R_{ki} = \) complete-case indicator; \( R_{ki} = 1 \) if \( V_{ki} \) is known or if \( T_{ki} \) is censored and \( R_{ki} = 0 \) otherwise

- **Auxiliary variables** \( A_{ki} \) can be used to predict whether the mark is missing and to predict the missing marks
  - E.g., \( A_{ki} = \) sequence information from a later sampled virus
- Model the relationship between \( A_{ki} \) and \( V_{ki} \) to predict \( V_{ki} \)
• $r_k(W_{ki}, \psi_k) = \text{parametric model for the probability of complete-case, where } \psi_k \text{ is a } q\text{-dimensional parameter}$

• The IPW estimator $\hat{\beta}^{ipw}(v)$ solves the estimating equation for $\beta$:

$$U_{ipw}(v, \beta, \hat{\psi}) = \sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} \int_{0}^{1} \int_{0}^{\tau} K_h(u - v)(Z_{ki}(t) - \tilde{Z}_k(t, \beta, \hat{\psi}_k)) \frac{R_{ki}}{\pi_k(Q_{ki}, \hat{\psi}_k)} N_{ki}(dt, du),$$

where

$$\tilde{Z}_k(t, \beta, \psi_k) = \tilde{S}_k^{(1)}(t, \beta, \psi_k)/\tilde{S}_k^{(0)}(t, \beta, \psi_k),$$

$$\tilde{S}_k^{(j)}(t, \beta, \psi_k) = n_k^{-1} \sum_{i=1}^{n_k} R_{ki}(\pi_k(Q_{ki}, \psi_k))^{-1} Y_{ki}(t) \exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^\otimes j$$
Augmented IPW Complete-Case Estimator

- $W_{ki} = (T_{ki}, Z_{ki}, A_{ki})$ and $w = (t, z, a)$

More efficient estimation can be achieved by incorporating the knowledge of the conditional mark distribution:

$$
\rho_k(w, v) = P(V_{ki} \leq v|\delta_{ki} = 1, W_{ki} = w)
= \frac{\int_0^v \lambda_k(t, u|z)g_k(a|t, u, z)du}{\int_0^1 \lambda_k(t, u|z)g_k(a|t, u, z)du},
$$

where $g_k(a|t, v, z) = P(A_{ki} = a|T_{ki} = t, V_{ki} = v, Z_{ki} = z, \delta_{ki} = 1)$

- Let $\hat{g}_k(a|t, u, z)$ be a parametric / semiparametric estimator of $g_k(a|t, u, z)$; then $\rho_k(w, v)$ can be estimated by

$$
\hat{\rho}^{ipw}_k (w, v) = \frac{\int_0^v \hat{\lambda}^{ipw}_k(t, u|z)\hat{g}_k(a|t, u, z)du}{\int_0^1 \hat{\lambda}^{ipw}_k(t, u|z)\hat{g}_k(a|t, u, z)du}
$$
• Assessed how VE against subtype CRF01_AE HIV-1 infection depends on a weighted Hamming distance (Nickle et al., 2007, *PLoS One*) of breakthrough HIV-1 sequences to the A244 reference sequence contained in the vaccine
  • Include published gp120 AA sites in contact with broadly neutralizing monoclonal antibodies

• $T =$ time to HIV-1 infection diagnosis with subtype CRF01_AE HIV-1
  • Infection with subtype B or unknown subtype treated as right-censoring

• 106 HIV-1 subtype CRF01_AE infected participants (42 vaccine, 64 placebo); 94 (37 vaccine, 57 placebo) with an observed mark
  • Between 2 and 13 HIV-1 sequences (total 1030 sequences) per infected participant
  • $V =$ participant-specific median distance
Figure: Boxplots of the marks/distances $V$ for the 94 HIV-1 CRF01_AE infected subjects in the Thai trial with an observed mark.
Vaccine Efficacy by gp120 HIV-1 Sequence Distance

Figure: IPW point and 95% interval estimates of $VE(v)$ for the Thai trial with bandwidths $h_1 = 0.5$, $h_2 = h = 0.3$


4. Extension of 3. for covariate-adjustment and modeling dropout (Benkeser, Carone, Gilbert, 2016, submitted, tMLE)

5. Cumulative incidence VE for a continuous mark genotype (Gilbert, Sun, and McKeage, 2008, Biostatistics)

6. Proportional hazards VE for a continuous mark genotype (Sun, Gilbert, and McKeage, 2009, Ann Stat; local partial likelihood and kernel smoothing)
