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

Session 1 (Halloran)	Introduction to Study Designs for Evaluating VE
Session 2 (Follmann)	Introduction to Vaccinology Assays and Immune Response
Session 3 (Gilbert)	Introduction to Frameworks for Assessing Surrogate Endpoints/Immunological Correlates of VE
Session 4 (Follmann)	Additional Study Designs for Evaluating VE
Session 5 (Gilbert)	Methods for Assessing Immunological Correlates of Risk and Optimal Surrogate Endpoints
Session 6 (Gilbert)	Effect Modifier Methods for Assessing Immunological Correlates of VE (Part I)
Session 7 (Gabriel)	Effect Modifier Methods for Assessing Immunological Correlates of VE (Part II)
Session 8 (Sachs)	Tutorial for the R Package <i>pseval</i> for Effect Modifier Methods for Assessing Immunological Correlates of VE
Session 9 (Gilbert)	Introduction to Sieve Analysis of Pathogen Sequences, for Assessing How VE Depends on Pathogen Genomics
Session 10 (Follmann)	Methods for VE and Sieve Analysis Accounting for Multiple Founders



- 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 = time from study entry (or post immunization series) until study endpoint through to time τ_1 (e.g., HIV-1 infection)
- t = fixed time point of interest $t < \tau_1$
- Discrete genotype-specific cumulative VE

$$V\!E^{\mathrm{cml/disc}}(t,j) = \left[1 - rac{P(T \le t, J = j | \mathsf{Vaccine})}{P(T \le t, J = j | \mathsf{Placebo})}
ight] imes 100\%, \ t \in [0, au_1]$$

• Continuous genetic distance-specific cumulative VE

$$V\!E^{ ext{cml/cont}}(t, v) = \left[1 - rac{P(T \leq t, V = v | ext{Vaccine})}{P(T \leq t, V = v | ext{Placebo})}
ight] imes 100\%, \ t \in [0, au_1]$$

- *J* = discrete genotype subgroup such as binary, unordered categorical, ordered categorical
- V = (approximately) continuous genetic distance to a vaccine sequence

PBG (VIDD FHCRC)

Sieve Analysis Methods

Cumulative VE Sieve Effect Tests

Fix t at the primary time point of interest

• $VE^{\text{cml/disc}}(t,j)$:

$$\begin{split} & H_0: V E^{\mathrm{cml/disc}}(t,j) \text{ constant in } j \\ & H_1^{mon}: V E^{\mathrm{cml/disc}}(t,j) \text{ decreases in } j \\ & H_1^{any}: V E^{\mathrm{cml/disc}}(t,j) \text{ has some differences in } j \end{split}$$

• $VE^{\text{cml/cont}}(t, v)$:

$$\begin{split} H_0: V E^{\mathrm{cml/cont}}(t,v) & \text{constant in } v \\ H_1^{mon}: V E^{\mathrm{cml/cont}}(t,v) & \text{decreases in } v \\ H_1^{any}: V E^{\mathrm{cml/cont}}(t,v) & \text{has some differences in } v \end{split}$$

A "sieve effect" is defined by H_1^{mon} or H_1^{any} being true (i.e., differential VE by pathogen genotype)

PBG (VIDD FHCRC)

Illustration: Cumulative $VE^{cml/disc}(t = 14, j)$ for 3-Level J^*



*Aalen-Johansen (1978, *Scand J Stat*) nonparametric MLE (Aalen, 1978, *Ann Stat*; Johansen, 1978, *SJS*); test for differential *VE* by Neafsey, Juraska et al. (2015, *NEJM*)

PBG (VIDD FHCRC)

Sieve Analysis Methods

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



Continuous Genetic Distance-Specific Cumulative VE at t = 14 Months

*Aalen-Johansen (1978, *Scand J Stat*) nonparametric MLE (Aalen, 1978, *Ann Stat*; Johansen, 1978, *SJS*); test for differential *VE* by Neafsey, Juraska et al. (2015, *NEJM*)

PBG (VIDD FHCRC)

Sieve Analysis Methods

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)

Instantaneous Genotype-Specific VE Parameters

- h(t,j) = Hazard of the malaria endpoint with discrete genotype j
- $\lambda(t, v) =$ 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

$$VE^{
m haz/cont}(t,v) = \left[1 - rac{\lambda(t,v|
m Vaccine)}{\lambda(t,v|
m Placebo)}
ight] imes 100\%$$

• Proportional hazards assumption: $VE^{haz/disc}(t,j) = VE^{haz/disc}(j)$ and $VE^{haz/cont}(t,v) = VE^{haz/cont}(v)$ for all $t \in [0, \tau_1]$

Illustration: Instantaneous $VE^{haz/disc}(j)$ for 3-Level J^*



Discrete Genotype–Specific Instantaneous VE to 14 Months

*Gilbert (2000, *Stat Med*): genotype-specific Cox model

PBG (VIDD FHCRC)

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



Continuous Genetic Distance-Specific Instantaneous VE to 14 Months

Genetic Distance to Vaccine Insert Sequence

*Juraska and Gilbert (2013, *Biometrics*): overall endpoint Cox model + semiparametric biased sampling model

PBG (VIDD FHCRC)

Sieve Analysis Methods

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

1 Sieve Analysis Via Cumulative and Instantaneous VE Parameters

@ 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

Ongoing Sieve Analysis Statistical Methods Research

- 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

$$VE^{\mathrm{cml/disc}}(t,j) = \left[1 - \frac{P(T \le t, J = j | \mathsf{Vaccine})}{P(T \le t, J = j | \mathsf{Placebo})}
ight] imes 100\%, \ t \in [0, au_1]$$

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

$$ar{Q}_j^z(t) \equiv P(ilde{T}=t,\Delta J=j|Z=z, ilde{T}>t-1) \ ar{Q}_{\cdot}^z(t) \equiv \sum_{i=1}^K ar{Q}_i^z(t)$$

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

PBG (VIDD FHCRC)

Sieve Analysis Methods

Cumulative Genotype-Specific VE: Aalen-Johansen NPMLE

• Aalen-Johansen estimator plugs in empirical estimates

$$\bar{Q}_{j,n}^{z}(t) = \frac{\text{No. type } j \text{ events at } t \text{ in group } z}{\text{No. at risk at } t-1 \text{ in group } z}$$
$$\widehat{P}(T \le t, J = j | Z = z) = \sum_{t'=1}^{t} \left[\bar{Q}_{j,n}^{z}(t') \prod_{s=1}^{t'-1} \{1 - \bar{Q}_{\cdot,n}^{z}(s)\} \right]$$

Limitations

- For consistency need random censoring (cannot depend on covariates)
- Efficient if no prognostic factors

Incorporating Covariates: TMLE

$$P(T \le t, J = j | Z = z) = E_W [P(T \le t, J = j | Z = z, W)]$$

=
$$\sum_w P(T \le t, J = j | Z = z, W = w) P(W = w | Z = z)$$

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

Mean Squared Error TMLE vs. Aalen-Johansen



Power of Wald Tests TMLE vs. Aalen-Johansen



Power of Wald Tests TMLE vs. Aalen-Johansen



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
- $\widehat{VE} = 31\%$, 95% CI 1% to 51%, p = 0.04 (Rerks-Ngarm et al., 2009, *NEJM*)



• Cox model (Lunn and McNeil, 1995, *Biometrics*) and Aalen-Johansen (1978) sieve analysis yielded the inference

$$VE^{cml/disc}(3.5, v=0) > VE^{cml/disc}(3.5, v=1)$$

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

• TMLE adjusting for rish behaviors, gender, age, gave a similar result with increased precision (Benkeser, Carone, Gilbert, 2016); next slide

TMLE Cumulative VE Sieve Results: RV144 Thai Trial

AA position 169 matched



AA position 169 mismatched

PBG (VIDD FHCRC)

- 1 Sieve Analysis Via Cumulative and Instantaneous VE Parameters
- Q 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

- Sun and Gilbert (2012, Scand J Stat)
- 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:

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

• Covariate-adjusted mark-specific vaccine VE:

$$\operatorname{VE}(t, v|z) = 1 - rac{\lambda_v(t, v|z)}{\lambda_p(t, v|z)},$$

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:

$$\lambda_k(t, \mathbf{v} | \mathbf{z}_{ki}(t)) = \lambda_{0k}(t, \mathbf{v}) \exp\left\{\beta(\mathbf{v})^T \mathbf{z}_{ki}(t)\right\}, k = 1, \dots, K$$

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

z = (z₁, z₂); z₁ = vaccine group indicator; z₂ other covariates; β₁(v) = coefficient corresponding to z₁

Mark-specific vaccine efficacy:

$$VE(v) = 1 - \exp(eta_1(v))$$

Completely observed competing risks data:

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

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

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

$$O_{ki} = \{X_{ki}, Z_{ki}, \delta_{ki}, R_{ki}, R_{ki}\delta_{ki}V_{ki}, \delta_{ki}A_{ki}\}, i = 1 \dots, n_k, k = 1, \dots, K,$$

 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}

Inverse Probability Weighted Complete-Case Estimator

- $r_k(W_{ki}, \psi_k)$ = parametric model for the probability of complete-case, where ψ_k is a *q*-dimensional parameter
- The IPW estimator $\hat{\beta}^{ipw}(v)$ solves the estimating equation for β :

$$U_{ipw}(v,\beta,\hat{\psi}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau \mathcal{K}_h(u-v) \left(Z_{ki}(t) - \tilde{Z}_k(t,\beta,\hat{\psi}_k) \right) \frac{R_{ki}}{\pi_k(Q_{ki},\hat{\psi}_k)} N_{ki}(dt,du),$$

where

$$\begin{split} \tilde{Z}_{k}(t,\beta,\psi_{k}) &= \tilde{S}_{k}^{(1)}(t,\beta,\psi_{k})/\tilde{S}_{k}^{(0)}(t,\beta,\psi_{k}), \\ \tilde{S}_{k}^{(j)}(t,\beta,\psi_{k}) &= n_{k}^{-1}\sum_{i=1}^{n_{k}}R_{ki}(\pi_{k}(Q_{ki},\psi_{k}))^{-1}Y_{ki}(t)\exp\{\beta^{T}Z_{ki}(t)\}Z_{ki}(t)^{\otimes j} \end{split}$$

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:

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

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

 Let ĝ_k(a|t, u, z) be a parametric / semiparametric estimator of g_k(a|t, u, z); then ρ_k(w, v) can be estimated by

$$\hat{\rho}_{k}^{ipw}(w,v) = \frac{\int_{0}^{v} \hat{\lambda}_{k}^{ipw}(t,u|z)\hat{g}_{k}(a|t,u,z) \, du}{\int_{0}^{1} \hat{\lambda}_{k}^{ipw}(t,u|z)\hat{g}_{k}(a|t,u,z) \, du}$$

Analysis of the RV144 Thai Trial

- 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_ 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

HIV-1 Sequence Distances to the Vaccine Sequence A244



Distances of HIV Envelope gp120 sequences to the A244 reference sequence (V)

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 $h_1 = 0.5$, $h_2 = h = 0.3$

Selected Literature on Sieve Analysis Methods

- Proportional hazards VE for a discrete genotype (Gilbert, 2000, 2001, Stat Med, Cox model)
- Extension of 1. accounting for missing data on genotypes (Hyun, Lee, and Sun, 2012, J Stat Plan Inference, AIPW)
- Cumulative incidence VE for a discrete genotype (Gilbert, 2000, 2001, Stat Med, Aalen-Johansen NPMLE)
- Extension of 3. for covariate-adjustment and modeling dropout (Benkeser, Carone, Gilbert, 2016, submitted, tMLE)
- Cumulative incidence VE for a continuous mark genotype (Gilbert, Sun, and McKeague, 2008, *Biostatistics*)
- Proportional hazards VE for a continuous mark genotype (Sun, Gilbert, and McKeague, 2009, Ann Stat; local partial likelihood and kernel smoothing)
- Extension of 6. for multivariate continuous mark genotypes (Sun and Gilbert, 2013, *Biostatistics*, local partial likelihood and kernel smoothing; Juraska and Gilbert, 2013, Biometrics, Cox model + semiparametric biased sampling model)
- Extension of 6. allowing missing data on genotypes (Sun and Gilbert, 2012, Scand J Stat, Gilbert and Sun, 2012, JRSS-B, add AIPW; Juraska and Gilbert, 2015, LIDA, add IPW)

PBG (VIDD FHCRC)