Case-Cohort Approach to Assessing Immunological Correlates of Risk, With Application to Vax004

Biostat 578A: Lecture 11

A manuscript pertinent to this talk is posted on the course webpage (JIDimmune.article.2005.pdf)

Outline

- Design of Vax004 for assessing immunological correlates of risk
- Methods: Case-cohort sampling design Cox proportional hazards model
- Application to Vax004

Assessing Antibodies as Correlates of Risk in Vax004

- **Secondary objective**: Assess if various *in vitro* measurements of antibody levels in vaccinees correlate with HIV infection rate
- 8 antibody assays that measure binding/neutralization of the MN or GNE8 HIV strains
  - ELISAs to measure antibody binding: gp120, V2, V3, CD4 blocking
  - Functional assay: Neutralization of MN HIV-1

Assessing Antibodies as Correlates of Risk in Vax004

- Sampling design
  - Specimens collected:
    - Month 0, 1, 6, 12, 18, 24, 30, 36 (troughs)
    - Month 0.5, 1.5, 6.5, 12.5, 18.5, 24.5, 30.5 (peaks)
  - Specimens assayed:
    - Random “subcohort" of 5% of all vaccinees (n=174, all time points)
      - n=163/11 in subcohort uninfected/infected
    - All infected vaccinees (n=239, last sample prior to infection)

The Cox Model with The Case-Cohort Sampling Design

- Cox proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) exp\left\{\beta_0^T Z(t)\right\}$$

  - $\lambda(t|Z) =$ conditional failure hazard given covariate history until time t
  - $\beta_0 =$ unknown vector-valued parameter
  - $\lambda_0(t) = \lambda(t|0) =$ unspecified baseline hazard function
    - Z are “expensive" covariates only measured on failures and subjects in the subcohort

Notation and Set-Up (Matches Kulich and Lin, 2004, JASA)

- $T =$ failure time (e.g., time to HIV infection diagnosis)
- $C =$ censoring time
- $X = min(T,C), \Delta = I(T \leq C)$
- $N(t) = I(X \leq t, \Delta = 1)$
- $Y(t) = I(X \geq t)$
- Cases are subjects with $\Delta = 1$
- Controls are subjects with $\Delta = 0$

Notation and Set-Up (Matches Kulich and Lin, 2004, JASA)

- Consider a cohort of n subjects, who are stratified by a variable V with K categories
- $\varepsilon =$ indicator of whether a subject is selected into the subcohort
    - $\alpha_k = Pr(\varepsilon = 1|V = k)$, where $\alpha_k > 0$
- $(X_{ki}, \Delta_{ki}, Z_{ki}(t), 0 \leq t \leq \tau, V_{ki}, \varepsilon_{ki} \equiv 1)$ observed for all subcohort subjects
- At least $(X_{ki}, \Delta_{ki} \equiv 1, Z_{ki}(X_{ki}))$ observed for all cases

Estimation of β_0

- With full data, β_0 would be estimated by the MPLE, defined as the root of the score function

$$U_F(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}_F(t,\beta)\}\, dN_i(t), \qquad (1)$$

where

$$\bar{Z}_F(t,\beta) = S_F^{(1)}(t,\beta)/S_F^{(0)}(t,\beta);$$

$$\begin{aligned} S_F^{(1)}(t,\beta) &= n^{-1}\sum_{i=1}^{n} Z_i(t) exp\{\beta^T Z_i(t)\} Y_i(t) \\ S_F^{(0)}(t,\beta) &= n^{-1}\sum_{i=1}^{n} exp\{\beta^T Z_i(t)\} Y_i(t) \end{aligned}$$

Estimation of β_0

- Due to missing data (1) cannot be calculated under the case-cohort design
- Many modified estimators have been proposed, all of which replace $\bar{Z}_F(t,\beta)$ with an approximation $\bar{Z}_C(t,\beta)$, so are roots of

$$U_C(\beta) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^{\tau} \{Z_{ki}(t) - \bar{Z}_C(t,\beta)\}\, dN_{ki}(t)$$

  - The double indices k, i reflect the stratification

Estimation of β_0

- The case-cohort at-risk average is defined as

$$\bar{Z}_C(t,\beta) \equiv S_C^{(1)}(t,\beta)/S_C^{(0)}(t,\beta),$$

where

$$\begin{aligned} S_C^{(1)}(t,\beta) &= n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_{ki}(t)Z_{ki}(t)exp\left\{\beta^T Z_{ki}(t)\right\}Y_{ki}(t) \\ S_C^{(0)}(t,\beta) &= n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_{ki}(t)exp\left\{\beta^T Z_{ki}(t)\right\}Y_{ki}(t) \end{aligned}$$

Estimation of β_0

- The potentially time-varying weight $\rho_{ki}(t)$ is set to zero for subjects with incomplete data, eliminating them from the estimation
- Cases and subjects in the subcohort have $\rho_{ki}(t) > 0$
  - Usually $\rho_{ki}(t)$ is set as the **inverse estimated sampling probability** (Using the same idea as the Weighted GEE methods of Robins, Rotnitzky, and Zhao, 1994, 1995)
- Different case-cohort estimators are formed by different choices of weights $\rho_{ki}(t)$
- Two classess of estimators (N and D), described next

N-estimators

- The subcohort is considered a sample from all study subjects regardless of failure status
  - The whole covariate history $Z(t)$ is used for all subcohort subjects
  - For cases not in the subcohort, only $Z(T_i)$ (the covariate at the failure time) is used
- Prentice (1986, Biometrika): $\rho_i(t) = \varepsilon_i/\alpha$ for $t < T_i$ and $\rho_i(T_i) = 1/\alpha$
- Self and Prentice (1988, Ann Stat): $\rho_i(t) = \varepsilon_i/\alpha$ for all t

N-estimators

- General stratified N-estimator
  - $\rho_{ki}(t) = \varepsilon_i/\widehat{\alpha}_k(t)$ for $t < T_{ki}$ and $\rho_{ki}(T_{ki}) = 1$
    - $\widehat{\alpha}_k(t)$ is a possibly time-varying estimator of α_k
    - α_k is known by design, but nonetheless estimating α_k provides greater efficiency for estimating β_0 (Robins, Rotnitzky, Zhao,1994)
    - A time-varying weight can be obtained by calculating the fraction of the sampled subjects among those at risk at a given time point (Barlow, 1994; Borgan et al., 2000, Estimator I)

D-estimators

- Weight cases by 1 throughout their entire at-risk period
- D-estimators treat cases and controls completely separately
  - α_k apply to controls only, so that α_k should be estimated using data only from controls
- Conditional on failure status, the D-estimator case-cohort design is similar to that of the case-control design whether or not the subcohort sampling is done retrospectively

D-estimators

- General D-estimator

$$\rho_{ki}(t) = \Delta_{ki} + (1 - \Delta_{ki})\varepsilon_{ki}/\widehat{\alpha}_k(t)$$

  - Borgan et al. (2000, Estimator II) obtained by setting

$$\widehat{\alpha}_k(t) = \sum_i^n \varepsilon_{ki}(1 - \Delta_{ki})Y_{ki}(t) / \sum_i^n (1 - \Delta_{ki})Y_{ki}(t),$$

    i.e., the proportion of the sampled controls among those who remain at risk at time t

  - Under “Computing", the course webpage includes R code for Borgan’s Estimator II with a time-independent expensive covariate of interest (contributed by Michal Kulich)

Main Distinctions between N- and D-Estimators

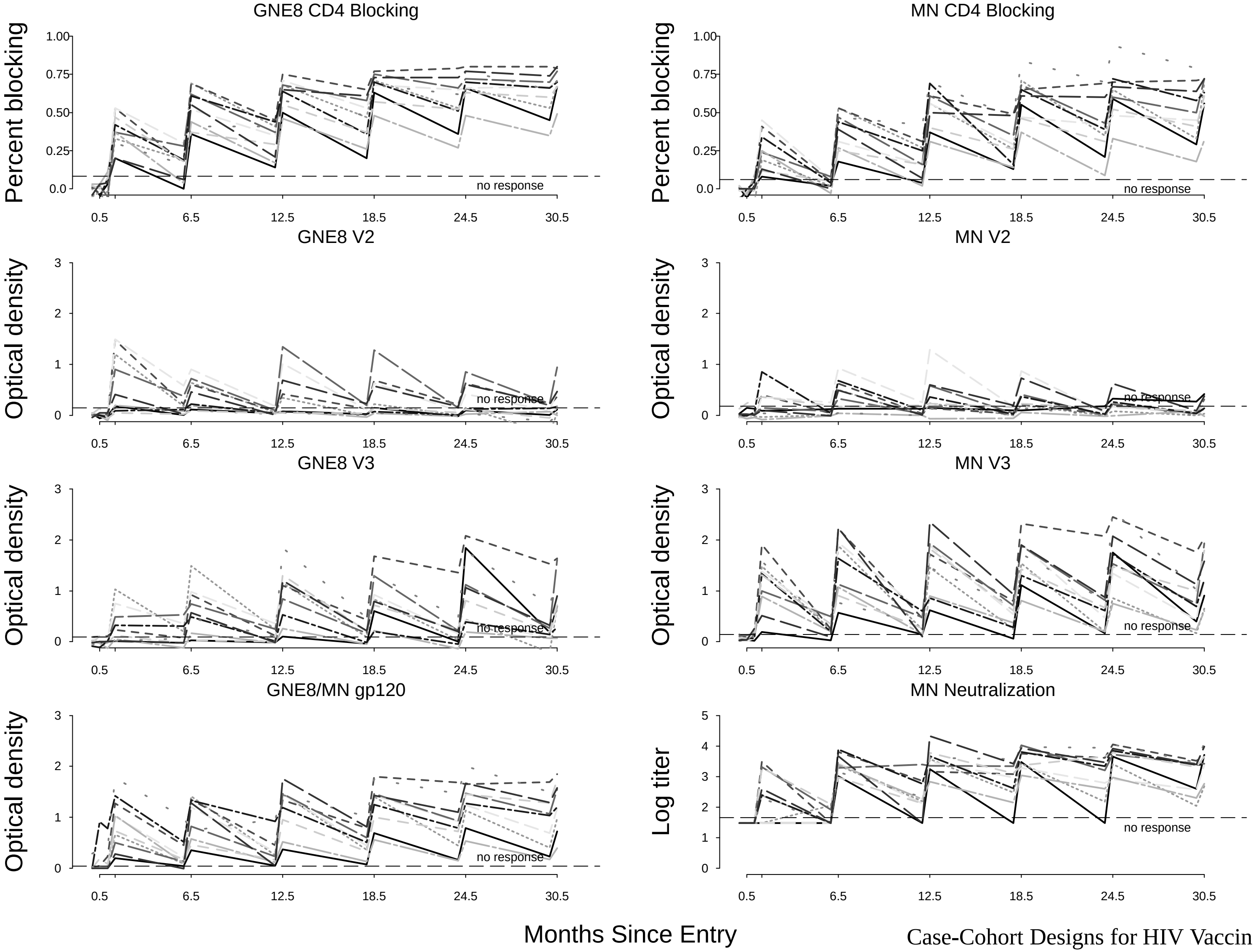
- D-estimators require data on the complete covariate histories of cases
- N-estimators only require data at the failure time for cases
    - For Vax004, the immune response in cases was only measured at the visit prior to infection, so N-estimators are valid while D-estimators are not valid

Main Distinctions between N- and D-Estimators

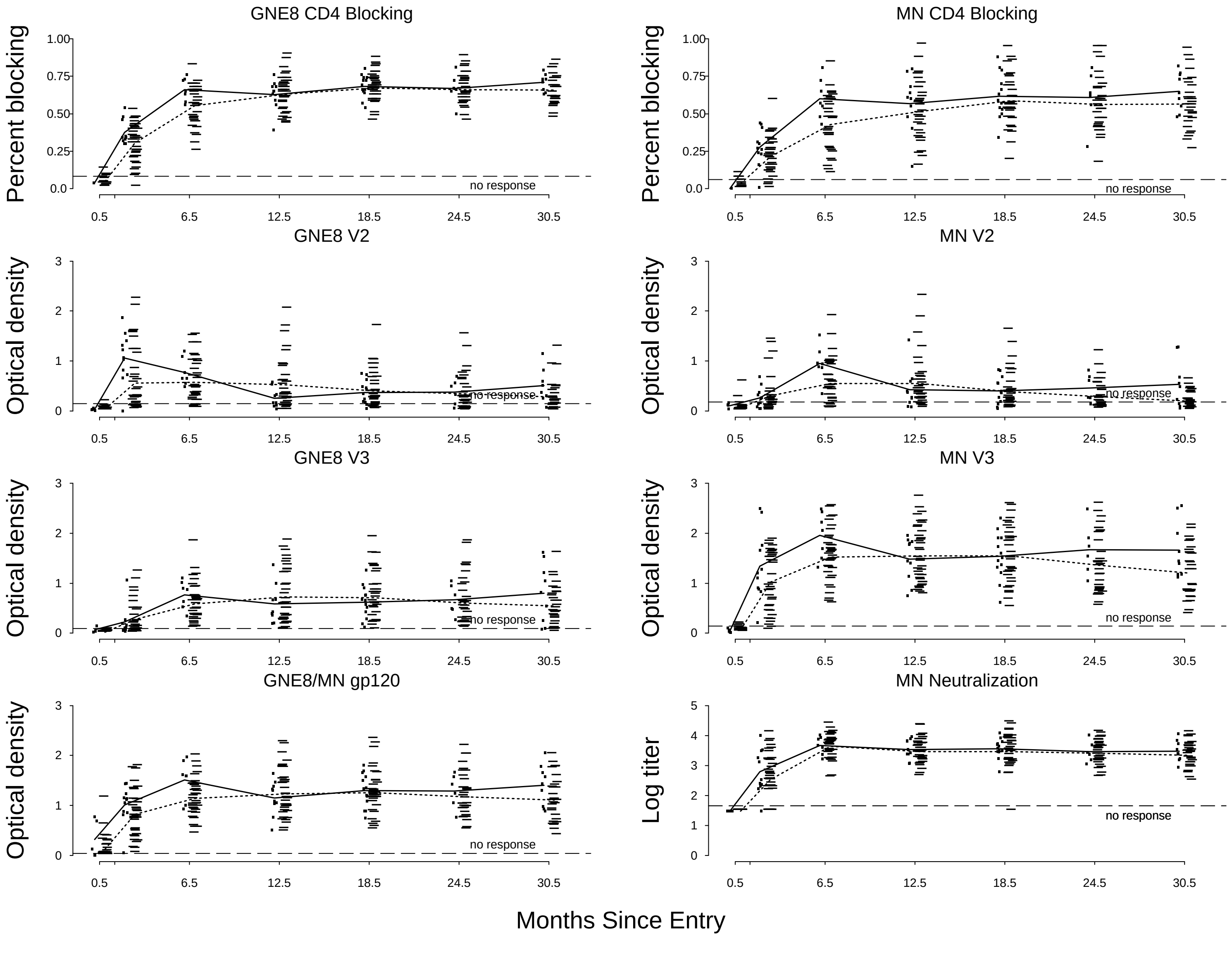
- For N-estimators, the sampling design is **specified in advance**, whereas for D-estimators, it can be **specified after the trial** (retrospectively)
  - D-estimators more flexible

Example: Application to Vax004

- Randomly selected subject-specific antibody profiles



Peak Antibody Levels of Vaccinees (Solid/dotted = Uninfected/infected)



Tests for Different Antibody Levels, Uninfected vs Infected Vaccinees

- Wei-Johnson (1985, Biometrika) tests linearly combine Wilcoxon statistics across the 7 time-points
- Overall/aggregate tests of whether peak antibody levels differ between infected (n=239) and uninfected (n=163) vaccinees

Antibody Variable	Wei-Johnson p-value
MN CD4	0.074
GNE8 CD4	0.0045
MN V2	0.13
GNE8 V2	0.18
MN V3	0.21
GNE8 V3	0.031
MN/GNE8 gp120	0.39
MN Neutralization	0.60

Results of Case-Cohort Cox Model Analysis

- Fit Prentice (1986) case-cohort Cox model, using $\widehat{\alpha} = 174/3598 = 0.0484$

Antibody	HR of HIV infection by Ab Quartile				P-value for	P-value for
variable	Q1	Q2	Q3	Q4	difference	trend
MN CD4	1.0	0.45	0.39	0.33	0.008	0.009
GNE8 CD4 Binding	1.0	0.46	0.37	0.30	0.026	0.013
MN V2	1.0	1.56	0.95	0.88	0.044	0.17
GNE8 V2	1.0	0.72	0.66	0.49	0.052	0.009
MN V3	1.0	0.88	0.59	0.84	0.22	0.39
GNE8 V3	1.0	0.45	0.53	0.40	0.011	0.030
MN/GNE8 gp120	1.0	0.96	0.69	0.68	0.30	0.096
MN Neutralization	1.0	0.52	0.42	0.46	0.080	0.088

Interpretation of Vax004 Results

- MN CD4 blocking, GNE8 CD4 blocking, GNE8 V2, GNE8 V3, MN Neutralization responses inversely correlated with HIV infection rate
- Estimated VE_S negative for low responses, $\approx$ zero for medium responses, positive for high responses
- Two possible explanations
  - High antibody levels cause protection and low antibody levels cause increased susceptibility [**Causation Hypothesis**]
  - Antibody levels mark individuals by their intrinsic risk of infection [**Association Hypothesis**]
- New methods needed to discriminate these
  - Addressed by Dean Follmann, covered in Lecture 12