Session 4 of Module 8: Evaluating an Immunological Correlate of Risk (Short Version, Expanded Slides at http://faculty.washington.edu/peterg/SISMID2013.html)

Peter Gilbert

Summer Institute in Statistics and Modeling in Infectious Diseases

U of W - July 15–17, 2013

Outline

1. CoR Methods: Case-cohort sampling design Cox proportional hazards model
   - Continuous time
2. Key issues
   - Sampling design
   - Measurement error
3. Improved analysis method (Breslow et al., 2009)

The Cox Model with a 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
  - β_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
- $\epsilon =$ indicator of whether a subject is selected into the subcohort
  - $\alpha_k = Pr(\epsilon = 1 | V = k)$, where $\alpha_k > 0$
- $(X_{ki}, \Delta_{ki}, Z_{ki}(t), 0 \leq t \leq \tau, V_{ki}, \epsilon_{ki} \equiv 1)$ 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

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

where

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

$$S_F^{(1)}(t, \beta) = n^{-1} \sum_{i=1}^{n} Z_i(t) exp \left\{ \beta^T Z_i(t) \right\} Y_i(t)$$

$$S_F^{(0)}(t, \beta) = n^{-1} \sum_{i=1}^{n} exp \left\{ \beta^T Z_i(t) \right\} Y_i(t)$$

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

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

  - 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 classes 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): $\rho_i(t) = \epsilon_i/\alpha$ for $t < T_i$ and $\rho_i(T_i) = 1/\alpha$
- Self and Prentice (1988, Ann Stat): $\rho_i(t) = \epsilon_i/\alpha$ for all t

N-estimators

- General stratified N-estimator
  - $\rho_{ki}(t) = \epsilon_i / \widehat{\alpha}_k(t)$ 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

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

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

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

    i.e., the proportion of the sampled controls among those who remain at risk at time t
  - the cch package in R (by Thomas Lumley and Norm Breslow) implements the case-cohort Cox model for N- and D-estimators (code for using cch to analyze a data set is provided at http://faculty.washington.edu/peterg/SISMID2013.html)

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

Gaps of Both N- and D- Estimators

Estimator	Does Not Need Full Covariate Histories in Cases	Allows Outcome-Dependent Sampling
N	Yes	No
D	No	Yes

- For time-dependent correlates, none of the partial-likelihood based methods are flexible on both points
- All of the methods require full covariate histories in controls
- Full likelihood-based methods can help (later)

Outline

1. CoR Methods: Case-cohort sampling design Cox proportional hazards model
   - Continuous time
2. Key issues
   - Sampling design
   - Measurement error
3. Improved analysis method (Breslow et al., 2009)

Some Sampling Questions to Consider Further

- Prospective or retrospective sampling?
- How much of the cohort to sample?
- Sampling design: Which subjects to sample?

Prospective or Retrospective Sampling?

Prospective sampling: Select a random sample for immunogenicity measurement **at baseline**

- Advantages of prospective sampling
  - Can estimate case incidence for groups with certain immune responses
  - Can study correlations of immune response with multiple study endpoints
  - Practicality: The lab will know what subjects to sample as early as possible, and there is one simple subcohort list

Prospective or Retrospective Sampling?

Retrospective sampling: At or after the final analysis, select a random sample of controls for immunogenicity measurement

- Advantages of retrospective sampling
  - Can match controls to cases to obtain balance on important covariates
    - E.g., balanced sampling on a prognostic factor gains efficiency (balanced sampling = equal number of subjects sampled within each level of the prognostic factor for cases and controls)
  - Can flexibly adapt the sampling design in response to the results of the trial
    - E.g., Suppose the results indicate an interaction effect, with VE $>>$ 0 in a subgroup and $VE \approx 0$ in other subgroups. Could over-sample controls in the 'interesting' subgroup.

Retrospective sampling may also sample controls at periodic intervals during the study follow-up period

Prospective or Retrospective Sampling?

For cases where there is one primary endpoint and it is not of major interest to estimate absolute case incidence, retrospective sampling may be typically preferred

How Many Controls to Sample?

- In prevention trials, for which the clinical event rate is low, it is very expensive and unnecessary to sample all of the controls
  - E.g., VaxGen trial: 368 HIV infected cases; 5035 controls
  - **Rule of thumb:** A $K:1$ Control:Case ratio achieves relative efficiency of $1 - \frac{1}{1+K}$ compared to complete sampling

K	Relative Efficiency
1	0.50
2	0.67
3	0.75
4	0.80
5	0.83
10	0.91

Which Controls to Sample?

Two-Phase Sampling

- **Phase I:** All N trial participants are classified into K strata on the basis of information known for everyone: N_k in stratum k; $N = \sum_{k=1}^{K} N_k$
- **Phase II:** For each k, $n_k \leq N_k$ subjects are sampled at random, without replacement from stratum k, and 'expensive' information (i.e., the immunological biomarker S) is measured for the resulting $n = \sum_{k=1}^{K} n_k$ subjects

Which Controls to Sample?

Principle: Well-powered CoR evaluation requires broad variability in the biomarker response and in the risk of the clinical endpoint

- Can improve efficiency by over-sampling the "most informative" subjects
  - Disease cases (usually sampled at 100%)
  - Rare or unusual immune responses; or rare covariate patterns believed to affect immune response (e.g., HLA subgroups)
- Baseline auxiliary data measured in everyone most valuable when they predict the missing data (i.e., the biomarker of interest)

Outline

1. CoR Methods: Case-cohort sampling design Cox proportional hazards model
   - Continuous time
2. Key issues
   - Sampling design
   - Measurement error
3. Improved analysis method (Breslow et al., 2009)

Measurement Error

Measurement error can reduce power to detect a CoR

Illustrative Example

- 'True' CoR $S^* \sim N(0,1)$
- 'Measured CoR' $S = S^* + \epsilon$, $\epsilon \sim N(0, \sigma^2)$
- Infection status Y generated from $\Phi(\alpha + \beta S^*)$

  with α set to give $P(Y = 1|S^* = 0) = 0.20$ and β set to give $P(Y = 1|S^* = 1) = 0.15$

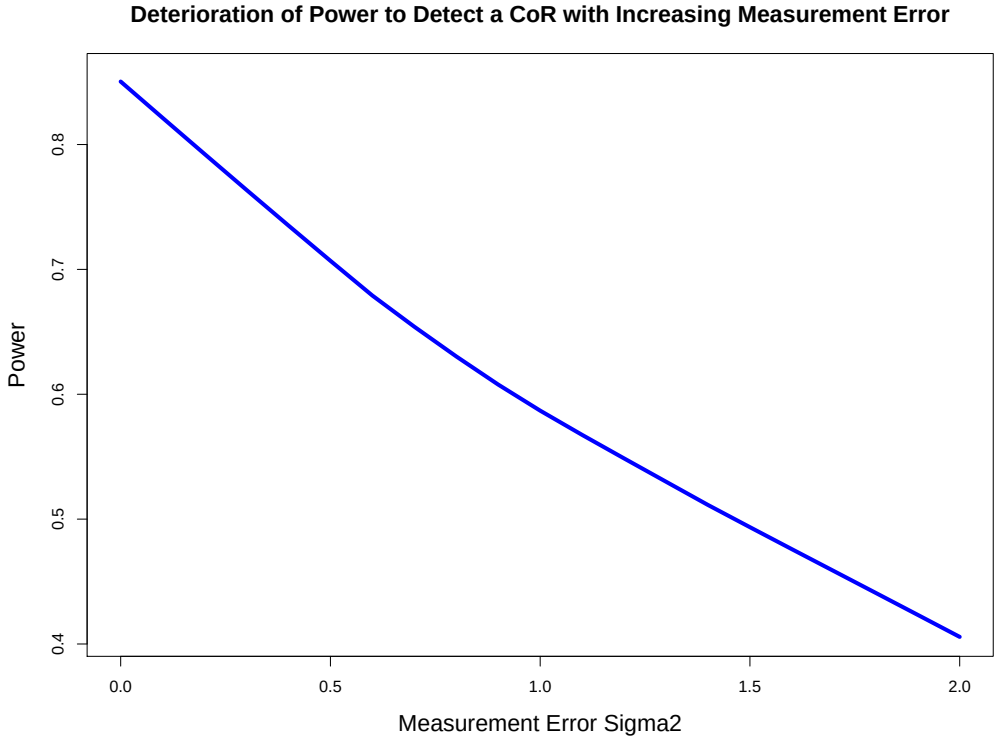
  σ^2 ranges from 0 to 2 (no-to-large measurement error)

Measurement Error Reduces Power

Simple Simulation Study

- Consider a study with $n = 500$ participants
- Consider power of a logistic regression model to detect an association between S and Y

Measurement Error Reduces Power



Outline

1. CoR Methods: Case-cohort sampling design Cox proportional hazards model
   - Continuous time
2. Key issues
   - Sampling design
   - Measurement error
3. Improved analysis method (Breslow et al., 2009)

Typical Correlates Assessments are Inefficient

- Broadly in epidemiology studies, biomarker-disease associations are commonly assessed ignoring much data collected in the study
- That is, only subjects with the biomarker measured (i.e., the Phase II sample) are included in the analysis
- Standard case-cohort analyses use inverse probability weighting of the subjects sampled in Phase 2, including all of the methods discussed so far
- These ubiquitously-used methods are implemented in the R package cch (Breslow and Lumley)

Typical Correlates Assessments are Inefficient

- Breslow et al.* urge epidemiologists to consider using the whole cohort in the analysis of case-cohort data
- Baseline data on demographics and potential confounders are typically collected in all subjects (the Phase I data measured in everyone)
- These Phase I data are most valuable when they predict "missing" data

*Breslow, Lumley et al. (2009, American Journal of Epidemiology; 2009, Statistical Biosciences)

How Leverage All of the Data?

- **Question:** How can we use the Phase I data to improve the assessment of CoRs?
- **Answer:** Adjust the sampling weights used in the conventional analyses
- The long version of these slides include 20 slides borrow from Professor Norman Breslow's Plenary Lecture at the World Congress of Epidemiology in Porto Alegre, Brazil, September 23, 2008.
- Long version of slides, plus R tutorial implementing the Breslow et al. (2009) method with the cch R package, are available at http://faculty.washington.edu/peterg/SISMID2013.html

Take Home Messages from Breslow et al., 2009

1. **Rule of thumb:** Obtain 'worthwhile' efficiency gain for the CoR assessment if baseline covariates can explain at least 40% of the variation in the immunological biomarker ($R^2 \geq 0.40$)
2. If interested in interactions (evaluation of whether a baseline covariate measured in everyone modifies the association of the biomarker and the clinical endpoint), can obtain worthwhile efficiency gain with a lower R^2
3. Even if no gain for the CoR assessment, will usually dramatically improve efficiency for assessing the associations of the Phase I covariates with outcome
4. Therefore it may often be the preferred method, and all practicing statisticians and epidemiologists should have the Breslow et al. method in their analytic toolkit
5. However, Breslow et al. (2009) currently only applies for a single immune response of interest measured at phase two, and does not handle a time-dependent immune response (serious practical limitationis that need more research to resolve)