Session 5 of Module 8: Methods for Assessing Immunological Correlates of Risk and Optimal Surrogate Endpoints

Peter Gilbert

Summer Institute in Statistics and Modeling in Infectious Diseases

U of W July 18–20, 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 *pseval* 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

Outline of Session 5

1. Traditional CoR methods: Inverse probability weighted Cox model
2. Key issues
   - Marker sampling design
   - Marker measurement error
3. Improved CoR methods (Breslow et al., 2009; Rose and van der Laan, 2011)
4. Estimated optimal surrogate (van der Laan, Price, Gilbert, 2016)

Prospective Cohort Study Sub-Sampling Design Nomenclature

- Terms used: case-cohort, nested case-control, 2-phase sampling
  - Case-cohort sampling originally meant taking a Bernoulli random sample of subjects at study entry for marker measurements (the "sub-cohort"), and also measuring the markers in all disease cases (Prentice, 1986, *Biometrika*)
  - Nested case-control sampling is Bernoulli or without replacement sampling done separately within disease cases and controls (retrospective sampling)
  - 2-phase sampling is the generalization of nested case-control sampling that samples within discrete levels of a covariate as well as within case and control strata (Breslow et al., 2009, *AJE*, *Stat Biosciences*)
  - Source of confusion: Some papers allow case-cohort to include retrospective sampling
- We restrict case-cohort to its original meaning

The Cox Model with a Sub-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 a random sub-sample
    - i.e., Z = immune response biomarkers, measured at fixed time τ post-randomization or at longitudinal visits

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 prospective cohort of N subjects, who are stratified by a variable V with K categories
- $\epsilon =$ indicator of whether a subject is selected for measurement of immune responses Z (and they are measured)
  - $\alpha_k = Pr(\epsilon = 1 | V = k)$, 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 marker 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 may be estimated by the MPLE, defined as the root of the score function

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

where

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

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

Estimation of β_0

- Due to missing data (1) cannot be calculated under the sub-sampling design
- Most estimators are based on pseudoscores parallel to (1), with $\bar{Z}_F(t, \beta)$ replaced with an approximation $\bar{Z}_C(t, \beta)$

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

  - The double indices k, i reflect the stratification

Estimation of β_0

- The marker sampled 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

- $\rho_{ki}(t)$ is set to zero for subjects with incomplete data, eliminating them from the estimation
- Cases and subjects in the marker 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 estimators are formed by different choices of weights $\rho_{ki}(t)$
- Two classes of estimators (case-cohort and 2-phase)

Example CoR Analysis: RV144 HIV-1 VE Trial

Haynes et al. (2012, *NEJM*) assessed in vaccine recipients the association of 6 immune response biomarkers measured at Week 26 with HIV-1 infection through 3.5 years

- **2-phase sampling design:** Measured Week 26 responses from all HIV-1 infected cases ($n = 41$) and from a stratified random sample of controls ($n = 205$ by gender $\times$# vaccinations $\times$ per-protocol)

Immune Response Variable	Est. HR (95% CI)	2-Sided P-value
IgA Magnitude-Breadth to Env	1.58 (1.07–2.32)	0.02
Avidity to A244 Strain	0.90 (0.55–1.46)	0.66
ADCC to 92TH023 Strain	0.92 (0.62–1.37)	0.67
Neutralization M-B to Env	1.46 (0.87–2.47)	0.15
IgG to gp70-V1V2 Env	0.57 (0.37–0.90)	0.014
CD4 T cell Magn to 92TH023	1.17 (0.83–1.65)	0.37

Borgan et al. (2000, *Lifetime Data Analysis*) Cox model estimator II

Case-cohort Estimators (Called N-estimators in Kulich and Lin, 2004)

- 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

Case-cohort 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)

Two-phase Sampling Estimators (Called D-estimators in Kulich and Lin, 2004)

- 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
- Nested case-control estimators are the special case with one covariate sampling stratum $K = 1$

Two-phase Sampling 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 Cox model for case-cohort (N-estimators) and 2-phase sampling (D-estimators) (code for using cch to analyze a data set is provided at http://faculty.washington.edu/peterg/SISMID2016.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
  - E.g., for the Vax004 HIV VE trial, the immune responses in cases were 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 (Prosp. case-cohort)	Yes	No
D (Retrosp. 2-phase)	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

Outline

1. Traditional CoR methods: Inverse probability weighted Cox model
2. Key issues
   - Marker sampling design
   - Marker measurement error
3. Improved CoR methods (Breslow et al., 2009; Rose and van der Laan, 2011)
4. Estimated optimal surrogate (van der Laan, Price, Gilbert, 2016)

Some Marker 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 case-cohort 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
  - Straightforward to descriptively study the distribution of the immune responses in the whole study population at-risk when the immune responses are measured
  - **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 2-phase sampling: At or after the final analysis, select a random sample of control subjects 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 effect modification, with VE $>>$ 0 in a subgroup and $VE \approx 0\%$ in other subgroups. Could over-sample controls in the 'interesting' subgroup.

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 referred

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
  - Vax004 trial vaccine recipients: 225 HIV infected cases; $\approx$ 3000 controls
  - RV144 trial vaccine recipients: 41 HIV infected cases; $\approx$ 7000 controls
  - **Rule of thumb:** Under the null hypothesis, 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

  - Simulations useful for studying the trade-offs of different K under alternative CoR hypotheses

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, and the 'expensive' immune response biomarkers Z are 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 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)
- Auxiliary Phase I variables measured in everyone are most valuable when they predict the missing data (i.e., the biomarker of interest)
- In general, optimal sampling obtained with sampling probabilities proportional to the cost-adjusted square-root variance of the efficient influence function (Gilbert, Yu, Rotnitzky, 2014, *Stat Med*)

Outline

1. Traditional CoR methods: Inverse probability weighted Cox model
2. Key issues
   - Marker sampling design
   - Marker measurement error
3. Improved CoR methods (Breslow et al., 2009; Rose and van der Laan, 2011)
4. Estimated optimal surrogate (van der Laan, Price, Gilbert, 2016)

Measurement Error Reduces 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 $P(Y = 1|S^* = 1) = 0.15$

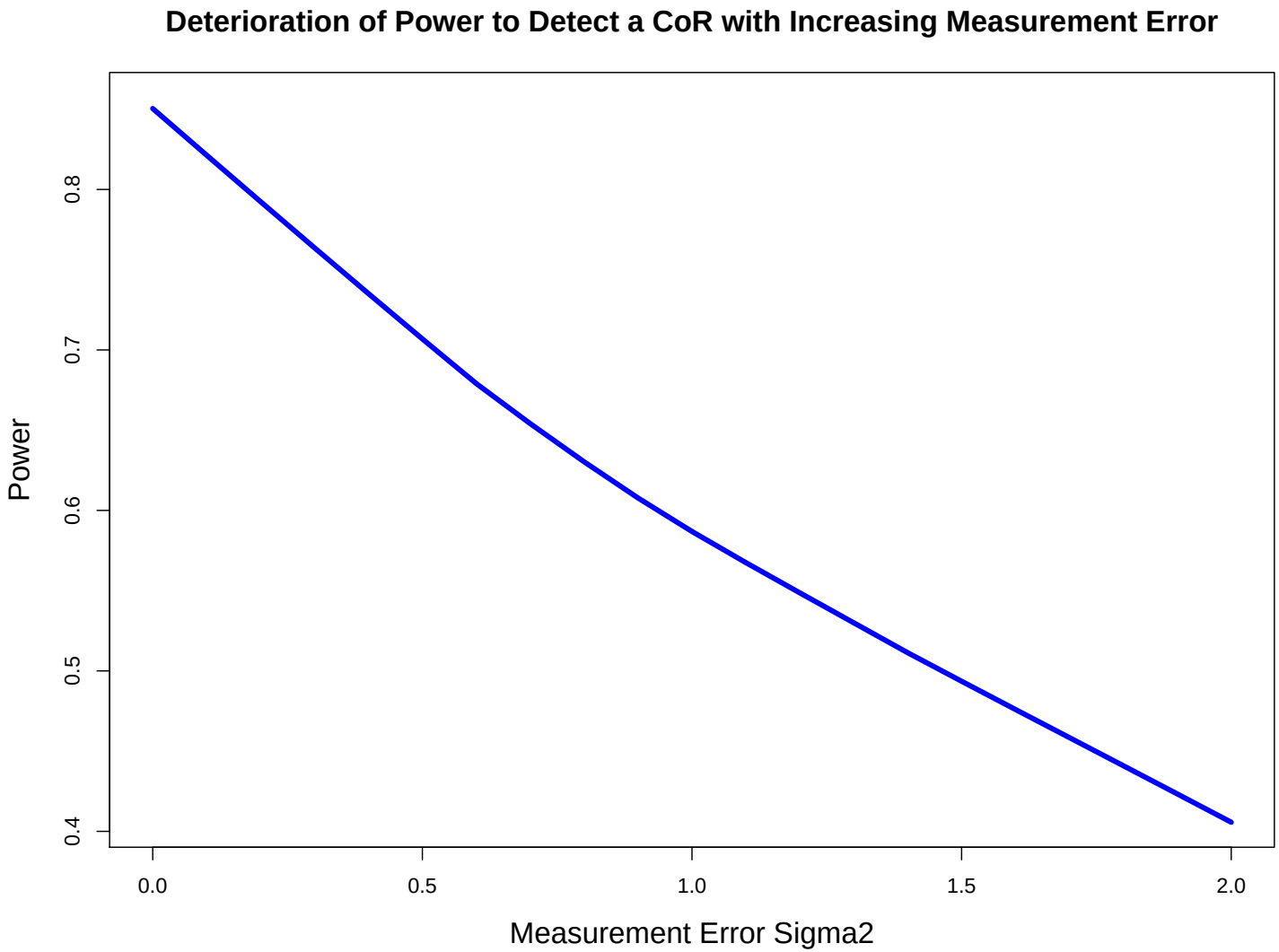
  σ^2 ranges from 0 to 2 (no-to-large measurement error)

Measurement Error Reduces Power to Detect a CoR

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

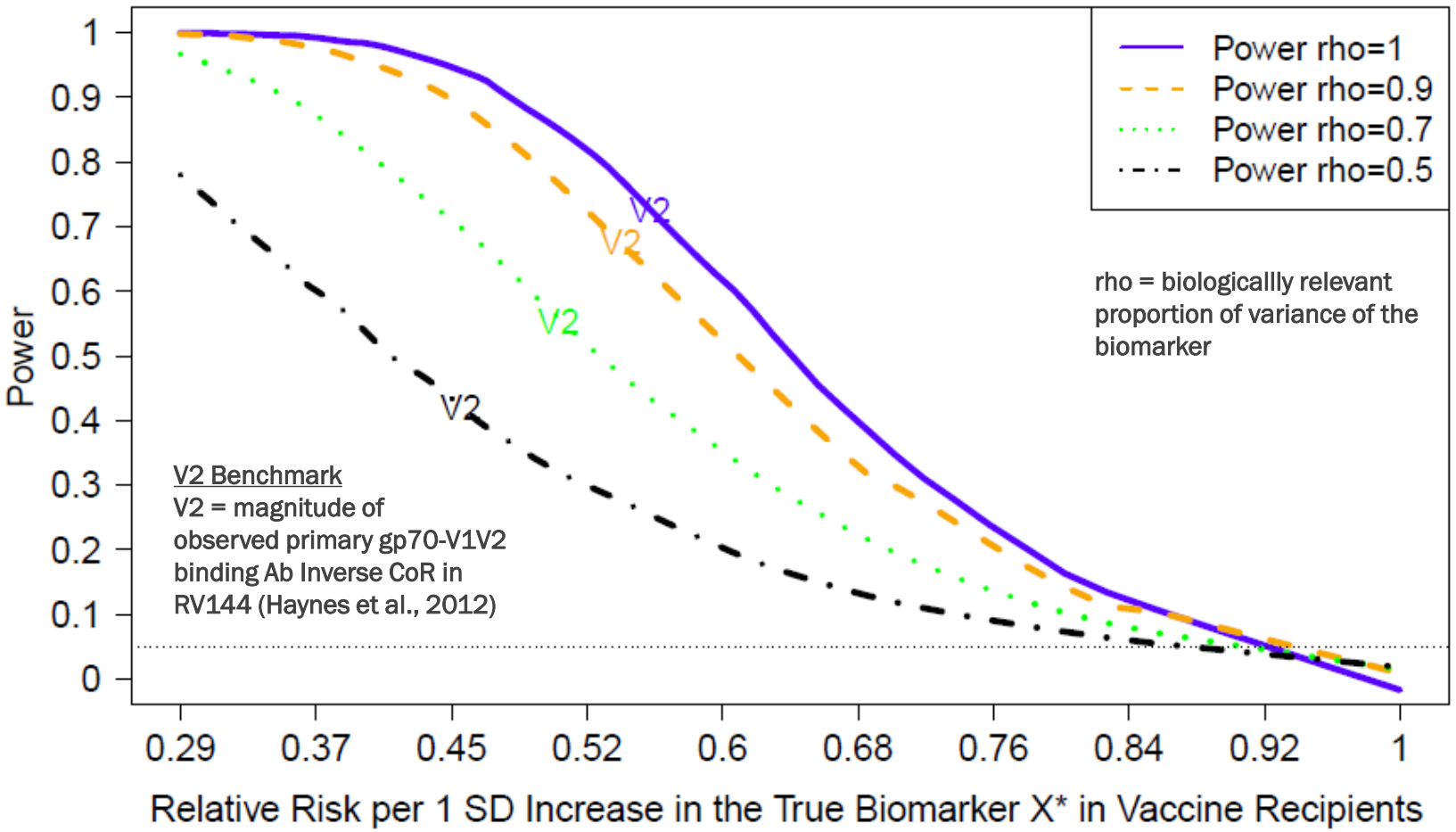
Measurement Error Reduces Power to Detect a CoR



Power Calculations for Assessing CoRs

- Ideally, the power/sample size calculations should explicitly account for measurement error in the assay
  - E.g., Gilbert, Janes, Huang (2016, *Stat Med*), implemented in the R package *CoRpower* posted at http://faculty.washington.edu/peterg/programs.html
  - E.g., specify $\rho \equiv \sigma^2/\sigma^2_{obs}$, the proportion of inter-vaccinee variability of the biomarker that is biologically relevant
  - **Rule of thumb:** $\rho =$relative efficiency for estimating a CoR odds ratio for the underlying perfect biomarker compared to the observed biomarker (McKeown-Eyssen, Tibshirani, 1994, *AJE*)
  - 'Noise' components of σ^2_{obs} may be estimated, especially from laboratory assay validation studies
    - Within-vaccinee variability of replicates
    - Between-vaccinee variability due to variability in the time from the last immunization to marker sampling

Power to Detect a CoR of HIV Infection in Vaccinees in HVTN 505 ($\alpha = 0.05$)



Method: 2-phase logistic regression (Holubkov and Breslow, 1997)

Outline

1. Traditional CoR methods: Inverse probability weighted Cox model
2. Key issues
   - Marker sampling design
   - Marker measurement error
3. Improved CoR methods (Breslow et al., 2009; Rose and van der Laan, 2011)
4. Estimated optimal surrogate (van der Laan, Price, Gilbert, 2016)

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 are included in the analysis
- Standard analyses use inverse probability weighting of the biomarker sampled subcohort, 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 statisticians/epidemiologists to consider using the whole cohort in the analysis of case-cohort/2-phase sampling 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, *AJE*, *Stat Biosciences*)

How to Leverage All of the Data?

- **Question:** How can we use the Phase I data to improve the assessment of CoRs?
- **One Answer:** One approach adjusts the sampling weights used in the standard analyses described above to obtain approximately efficient estimators (e.g., Breslow et al., 2009, *AJE*, *Stat Biosciences*)

Some Lessons Learned from Breslow et al. (2009)

1. 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 practicioners should have methods accounting for all of the data in their analytic toolkit
5. Additional research needed to make these more-efficient methods work well for multivariate markers and for time-dependent markers

How to Leverage All of the Data?

- **Question:** How can we use the Phase I data to improve the assessment of CoRs?
- **Another Answer:** Use an efficient and double-robust method: Inverse probability of censoring weighted targeted minimum loss based estimation (IPCW-TMLE) (Rose and Van der Laan, 2011, *Int J Biost*)

Right-Censored Data Structure for Fixed Follow-up Time t

- $V =$ Phase I information: Covariates (Z, V_0), $\tilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$, $Y^* = I(\tilde{T} \leq t)\Delta$, Phase II sampling probability ϵ
- $S = (A, W) =$ Phase II information: Immune response biomarkers measured at τ
  - Focus on the marker A of interest; $W =$ all other markers
  - Repeat the analysis taking each element of W as A

IPCW-TMLE: Target Parameters for Inference (Binary Marker)

Full data structure $X = (V, S) = (Z, V_0, \tilde{T}, \Delta, Y^*, A, W)$

General target parameters

- $P_{X,0}$ = true probability distribution of X
- M^F = statistical model for $P_{X,0}$
- $\Psi^F : \mathcal{M}^F \to R^d$ = target parameter of the full-data distribution
- $\psi_0^F = \Psi^F(P_{X,0})$ = target parameter of the true probability distribution of X

Causal risk target parameters for a binary marker A

$$\psi^F_{RD,0} = E_{X,0}\left[E_{X,0}(Y|A=1,W) - E_{X,0}(Y|A=0,W)\right] \quad (2)$$

$$\psi^F_{RR,0} = \frac{E_{X,0}\left[E_{X,0}(Y|A=1,W)\right]}{E_{X,0}\left[E_{X,0}(Y|A=0,W)\right]}$$

in each case with M^F nonparametric

IPCW-TMLE: Target Parameters for A Quantitative

- Idea from Alex Luedtke
  - Make inferences about

$$\max_{l<u}\Psi^F_{RD}(P_{X,0};l,u) = \max_{l<u}\{E[Y|A \geq u] - E[Y|A \leq l]\}$$

and

$$\max_{l<u}\Psi^F_{RR}(P_{X,0};l,u) = \max_{l<u}\left\{\frac{E[Y|A \leq l]}{E[Y|A \geq u]}\right\}$$

subject to a constraint on l and u such as that mentioned above

  - Assesses whether any trichotomization of the marker A yields a significant CoR, with the inference formally accounting for the searching for the best-discriminating cut-points

IPCW-TMLE: Data-Adaptive Target Parameters for A Quantitative

- Following Van der Laan, Hubbard, and Pajouh (2013), define data-adaptive causal contrasts using K-fold cross-validation
- Based on the first $K-1$ data pieces, define two cut-points $l_1 < u_1$ for A that maximize the IPCW-TMLE of $|\Psi^F_{RD}(P_{X,0}; l_1, u_1)|$, under a constraint such as $\geq 5\%$ cases with $A < l_1$ and with $A > u_1$
- Obtain IPCW-TMLE of $\Psi^F_{RD}(P_{X,0}; l_1, u_1)$ from withheld K^{th} piece
- Repeat for each set of $K-1$ data pieces with the K^{th} piece withheld, yielding K maximizing cutpoints $(l_1, u_1) \cdots (l_K, u_K)$ and K corresponding ICPW-TMLE estimators on withheld data sets
- Define the data-adaptive causal risk difference parameter as

$$\sum_{k=1}^{K} \Psi^F_{RD}(P_{X,0}; l_k, u_k),$$

estimated by the average of the IPCW-TMLE estimates

Implementation to Obtain the IPCW-TMLE of $E[Y|A]$

Observed data i.i.d. copies of $O = (Z, V_0, \tilde{T}, \Delta, Y^*, \epsilon, \epsilon A, \epsilon W)$

- If the full data X were available, then existing TMLE procedures could be used
- True target parameter $P_{X,0}$ defined wrt a specified full-data loss function $L^F(P_X)(X)$: $P_{X,0} = \text{argmin}_{P_X \in M^F} E_0 L(P_X)(X)$
    - **TMLE Step 1:** Construct an initial estimator $P^0_{X,n}$ of $P_{X,0}$
    - **TMLE Step 2:** Bias-correct $P^0_{X,n}$ through an iterative algorithm to yield $P^*_{X,n}$, making the empirical average of the full-data efficient influence curve at $P^*_{X,n}$ equaling zero, hence yielding an efficient estimator

Implementation to Obtain the IPCW-TMLE of $E[Y|A]$

- IPCW-TMLE proceeds in the same way, except in each step the following IPCW-loss function is used in place of the full-data loss function, where $\Pi_n(V)$ is a nonparametric or TMLE estimator of the marker sampling probability $\Pi_0(V) = P(\epsilon = 1|V)$

$$L(P_X)(O) \equiv \frac{\epsilon}{\Pi_n(V)} L^F(P_X)(X)$$

- Step 1 (initial estimation of $P_{X,0}$) can be maximally flexible and robust by using 2 or 3 superlearners for each element of $P_{X,0}$

1. Sampling probability estimator $\Pi_n(V)$
2. Conditional risk $E_{X,0}(Y|A,W)$
3. "Exposure mechanism" $g_0(a|W) \equiv P_{X,0}(A = a|W)$

Implementation to Obtain the IPCW-TMLE of $E[Y|A]$

Properties of IPCW-TMLE

- $\Pi_n(V)$ guaranteed consistent for $\Pi_0(V)$ if all the marker missingness is by design
- **Double-robustness property:** IPCW-TMLE is consistent even if the superlearner inconsistently estimates one (but not both) of $E_{X,0}(Y|A,W)$ and $g_0(a|W)$
- Consistent estimation of both terms implies the IPCW-TMLE is asymptotically efficient

Implementation to Obtain the IPCW-TMLE of $E[Y|A]$

- Ordinary superlearner (van der Laan, Polley, and Hubbard, 2007) may be used to estimate each piece $\Pi_0(V)$, $E_{X,0}(Y|A,W)$, $g_0(a|W)$, e.g., implemented with the Superlearner R package, using learners that allow specification of subject-specific weights $\epsilon_i/\Pi_n(V_i)$
- If there is substantial happenstance missingness of markers, then the missing at random assumption may fail
  - In this setting the superlearner for $\Pi_n(V)$ may be helpful
  - Neugebauer et al. (2013) demonstrated in marginal structural models that replacing a standard strategy of logistic regression modeling of the propensity score with superlearner reduced bias and improved efficiency

Outline

1. Traditional CoR methods: Inverse probability weighted Cox model
2. Key issues
   - Marker sampling design
   - Marker measurement error
3. Improved CoR methods (Breslow et al., 2009; Rose and van der Laan, 2011)
4. Estimated optimal surrogate (van der Laan, Price, Gilbert, 2016)

Introduction to an Optimal Surrogate

- **Goal:** Determine a surrogate outcome for a long-term outcome so that future randomized or observational studies can restrict themselves to only collecting the surrogate outcome
- Data from a clinical trial for developing a surrogate: n iid observations of $O = (W, Z, S, Y)$
  - W = Vector of baseline covariates
  - Z = Treatment assignment (e.g., 1=vaccine; 0=placebo)
  - S = Vector of response variables/markers at an intermediate time point τ
  - Y = Outcome of interest at a final time point after τ (binary or quantitative)
- Assume Z is randomized conditional on W

Introduction to an Optimal Surrogate

- Define an **optimal surrogate** for the current trial as the function of the data (W, Z, S) collected by the intermediate time point τ that optimally predicts the final outcome Y
    - A true parameter that we estimate with a targeted super-learner
- **Goal:** Use the estimated optimal surrogate in **future clinical trials** for estimation and testing of a mean contrast treatment effect on Y
    - Tackles the **transportability problem** of inferring the causal treatment effect in a new trial without measuring clinical endpoints Y (e.g., addressed by Pearl and Bareinboim, 2011, 2012)

Optimal Surrogate Framework vs. Other Frameworks

- **vs. controlled/natural effects and VE curve frameworks:** Departs by being based on average causal effects identified from standard assumptions in randomized trials
- **vs. Prentice/valid replacement endpoint framework:** Aligns in that the optimal surrogate satisfies the **Prentice definition**
  - Partially aligns with the **Prentice criteria**
    - The best optimal surrogate will have treatment and candidate surrogate highly predictive of Y, similar to Prentice criteria 1 and 2
    - The framework posits a conditional mean version of Prentice criterion 3 for licensing correct inferences on Y in a new trial
    - It handles equally well the general case where S varies or is constant in the placebo group
- **vs. meta-analysis framework:** Aligns in its objective of inference on the clinical treatment effect in a future study without collecting Y in that study (Gail et al., 2000, *Biostatistics*)
  - Departs in being based on a single (or few) trials and different transportability assumptions

Optimal Surrogate Framework

- Departs from all previous frameworks by defining the optimal surrogate as an unknown target parameter
  - Predicted values from the estimated optimal surrogate are used as the actual surrogate endpoint
  - In large samples this resulting surrogate must satisfy the Prentice definition (under the standard assumptions of an RCT)
- New approach in treating the surrogate endpoint problem as a supervised statistical learning problem
  - Previous methods evaluate a pre-selected univariable or low-dimensional vector candidate surrogate
  - The optimal surrogate approach is robust in that asymptotically consistent hypothesis tests and confidence intervals for the clinical treatment effects in the current and future trials are obtained without parametric modeling assumptions

Statistical Formulation of Estimation of an Optimal Surrogate

Observed data: iid copies $O = (W, Z, S, Y) \sim P_0$

- W = vector of baseline covariates
- Z = binary treatment assigned at baseline
- S = vector of intermediate outcomes measured at a fixed time point τ
- Y = final univariate outcome measured at a later final time point

- Potential outcomes (S_1, S_0) and (Y_1, Y_0) under treatment assignment $Z = 1$ and $Z = 0$
- Treatment Z is randomized conditional on W

A Nonparametric Approach

- $X = (W, S_0, S_1, Y_0, Y_1)$ = full-data structure with distribution $P_{X,0}$
- $O = (W, Z, S, Y)$ = observed data with distribution P_0 determined by $P_{X,0}$ and $g_0(z \mid X) = g_0(z \mid W)$
- The statistical model $\mathcal{M}$ for P_0 makes at most some assumptions about g_0
  - Known in a randomized trial
- $\mathcal{M}$ puts no assumptions on the marginal distribution of W nor on the conditional distribution of (S, Y) given A, W

Candidate Surrogate Outcomes

- Any real-valued function $(W, A, S) \rightarrow \psi(W, A, S) \in \mathbb{R}$ is a candidate surrogate, representing a measurement one can collect by time τ
- **Question:** How to define a good surrogate in terms of the true data distribution P_0?
- **Starting point:** We would like the surrogate $S^{\psi} \equiv \psi(W, A, S)$ to be valid in the actual study, according to the Prentice definition:

$$E_0(Y_1 - Y_0) = 0 \quad \text{if and only if} \quad E_0(S_1^{\psi} - S_0^{\psi}) = 0,$$

where $S_z^{\psi} = \psi(W, z, S_z)$, for $z \in \{0, 1\}$

  - Guarantees that an α-level test for $H_0^{\psi} : E_0(S_1^{\psi} - S_0^{\psi}) = 0$ is also an α-level test for $H_0 : E_0(Y_1 - Y_0) = 0$

Optimal Surrogate Outcome

- Criterion for ranking valid surrogates and defining a P_0-optimal surrogate: full-data mean squared error

$$\psi \rightarrow MSE_{P_{X,0}}(\psi) \equiv \sum_z E_{P_{X,0}} \left\{ g_0(z \mid W)(Y_z - \psi(W, z, S_z))^2 \right\}$$

- Goal is to minimize the weighted mean square prediction error for predicting Y_z across $z \in \{0, 1\}$
- Given a class $\mathbf{\Psi}$ of possible surrogate functions $\psi()$, the P_0-optimal surrogate in this class is defined as

$$\psi_0^F = \arg \min_{\psi \in \mathbf{\Psi}} MSE_{P_{X,0}}(\psi)$$

Optimal Surrogate Outcome

Theorem 1.
The minimizer of $\psi \to MSE_{P_{X,0}}(\psi)$ over all functions $(W, A, S) \to \psi(W, A, S)$ is:

$$\bar{S}_0 = \psi_0(W, Z, S) \equiv E_0(Y \mid W, Z, S)$$

Potential outcomes of this P_0-optimal surrogate: $\bar{S}_{0,z} = E_0(Y_z \mid W, S_z)$, $z \in \{0, 1\}$ and

$$E_{P_0}(\bar{S}_{0,z} \mid W) = E_{P_0}(Y_z \mid W)$$

- **Implication:** Under P_0, a 95% confidence interval for the causal effect of treatment on the P_0-optimal surrogate is also a 95% confidence interval for the causal effect of treatment on Y

Conditions for a New Study P Under Which the P_0-Optimal Surrogate is also the P-Optimal Surrogate

Theorem 2.
Consider a new study with iid observations $O^* = (W^*, Z^*, S^*, Y^*)$ with distribution P, where Z^* is randomized conditional on W^*

- **Transportability assumption:** $E[Y^*|W^* = w, Z^* = z, S^* = s] = E[Y|W = w, Z = z, S = s]$ for all (w, z, s) in a support of (W^*, Z^*, S^*)
- **Support assumption:** A support of (W^*, Z^*, S^*) is contained in a support of (W, Z, S)

Result: The P_0-optimal surrogate equals the P-optimal surrogate: for all (w, z, s) in a support of (W^*, Z^*, S^*),

$$E_P(Y^* \mid W^* = w, Z^* = z, S^* = s) = E_{P_0}(Y \mid W = w, Z = z, S = s)$$

and

$$E_P(Y^* \mid W^* = w, Z^* = z, S^* = s) = E_P(Y_z^* \mid W^* = w, S_z^* = s)$$

Transportability Theorem Under a Prentice Criterion 3: Application to a New Treatment $Z^* \neq Z$

- If the new study considers a new treatment $Z^* \neq Z$, then generally the transportability theorem will not apply, because $E[Y^*|W^* = w, Z^* = z, S^* = s] \neq E[Y|W = w, Z = z, S = s]$

Theorem 3.

- **Transportability and Support assumptions:** Same as in Theorem 2
- **Prentice criterion 3 assumption for both settings:** $E[Y^*|W^*, Z^*, S^*] = E[Y^*|W^*, S^*]$ and $E[Y|W, Z, S] = E[Y|W, S]$

Result: The P-optimal surrogate equals the P_0-optimal surrogate:

$$E_P(Y^* \mid W^* = w, Z^* = z, S^* = s) = E_{P_0}(Y \mid W = w, Z = z, S = s)$$
$$E_P(Y^* \mid W^* = w, Z^* = z, S^* = s) = E_P(Y_z^* \mid W^* = w, S_z^* = s)$$
$$E_{P_0}(Y_z \mid W = w, S_z = s) \;\&\; E_P(Y_z^* \mid W^* = w, S_z^* = s) \text{ constant in } z$$

Super-learning of the P_0-optimal surrogate

- Estimation of the P_0-optimal surrogate is a standard prediction problem
- Estimate $E_0(Y \mid W, Z, S)$ by a minimizer of the risk of a loss
  - E.g., for Y binary, use log-likelihood loss

$$L(\psi)(O) = -\{Y \log \psi(W, A, S) + (1 - Y) \log(1 - \psi(W, A, S))\}$$

- Loss-based super-learning*: yields an optimal estimator among any given class of candidate estimators
  - Oracle inequality for the cross-validation selector: the estimator is asymptotically at least as good as any candidate in the set of candidate estimators

*van der Laan, Polley, and Hubbard (2007); van der Laan and Rose (2011) textbook

Dengue Phase 3 Trial Example

- Two randomized, double-blinded, placebo-controlled, multicenter, Phase 3 trials of a recombinant, live, attenuated, tetravalent dengue vaccine (CYD-TDV)
  - **CYD14:** Asia-Pacific region (Capeding, et al., 2014, *The Lancet*)
  - **CYD15:** Latin America (Villar et al, 2015, *NEJM*)

Trial Designs

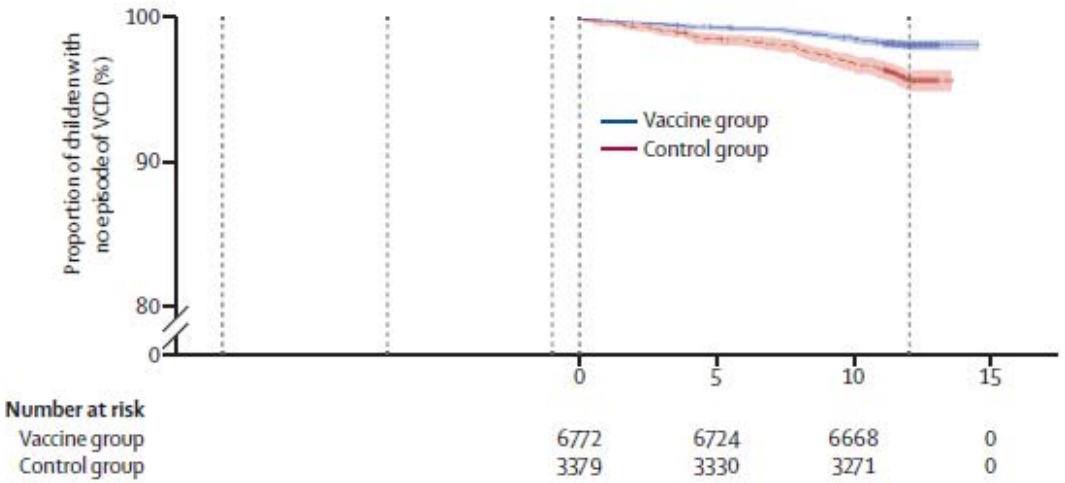
- 2:1 randomization to vaccine:placebo
- Immunizations at months 0, 6, 12
- Primary follow-up from Month 13 to Month 25 (active phase of follow-up)
- Primary endpoint: Symptomatic, virologically confirmed dengue (VCD)

Results on Vaccine Efficacy (Estimates from a Proportional Hazards Model)

CYD14: $\widehat{VE} = 56.5\%$ (95% CI 43.8–66.4)

CYD15: $\widehat{VE} = 64.7\%$ (95% CI 58.7–69.8)

CYD14 Trial (Capeding et al., 2014, *The Lancet*)



CYD15 Trial (Villar et al., 2015, *NEJM*)

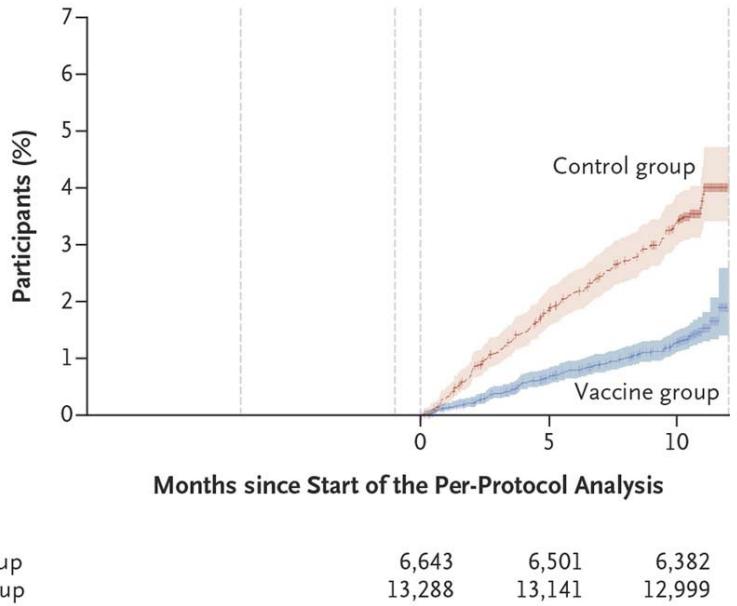


Illustration of Estimated Optimal Surrogate Approach

Analysis carried out by Brenda Price

- Based on pseudo CYD14 and CDY15 simulated data sets
- Treat CYD14 as the current trial; CYD15 as the future trial

Notation and Variables

- Z = Vaccination status (1=vaccine; 0=placebo)
- Y = Disease outcome (1=VCD endpoint between Month 13 and 25; 0 = no VCD endpoint by Month 25)
- W = Baseline covariates: age, sex, baseline PRNT_{50} neutralization titers to the 4 serotypes in the vaccine
- S = Month 13 PRNT_{50} neutralization titers to the 4 serotypes in the vaccine

Illustration of Estimated Optimal Surrogate Approach

Objectives

1. Estimate the P_0-optimal surrogate via targeted super-learner in CYD14, yielding $\psi_n^{TMLE}(W, A, S)$
2. Estimate VE^* in CYD15 based on the estimated optimal surrogate from CYD14 without using the CYD15 outcome data Y

Estimates of Dengue Risks and VEs in CYD14 and CYD15

1. Estimate the P_0-optimal surrogate via targeted super-learner in CYD14, yielding $\psi_n^{TMLE}(W, A, S)$. Obtain:

$$\begin{aligned} \widehat{E}[Y|Z=z] &= \frac{1}{n_z}\sum_{i=1}^{n_z} I(Z_i=z)\psi_n^{TMLE}(W_i, Z_i=z, S_i),\ z=0,1 \\ \widehat{VE} &= 1-\widehat{E}[Y|Z=1]/\widehat{E}[Y|Z=0] \end{aligned}$$

2. Estimate VE^* in CYD15 based on the estimated optimal surrogate from CYD14 without using the CYD15 outcome data Y

$$\begin{aligned} \widehat{E}[Y^*|Z^*=z] &= \frac{1}{n_z^*}\sum_{i=1}^{n_z^*} I(Z_i^*=z)\psi_n^{TMLE}(W_i^*, Z_i^*=z, S_i^*),\ z=0,1 \\ \widehat{VE}^* &= 1-\widehat{E}[Y^*|Z^*=1]/\widehat{E}[Y^*|Z^*=0] \end{aligned}$$

Wald 95% CIs based on influence functions and the delta method

Super-learner to Estimate the Optimal Surrogate

- Use the MSE loss function for the superlearner cross-validation selector (matched to the optimality criterion for a surrogate)

Table: Input Variables for the Learning Algorithms

Input Variables	
W:	Baseline demographics age (range 2–14 years), sex
W:	Baseline titers to the 4 serotypes inside the CYD-TDV vaccine, min and max of the 4 titers, interactions with age
S:	Month 13 titers to the 4 serotypes inside the CYD-TDV vaccine, min and max of the 4 titers, interactions with age

Super-learner to Estimate the Optimal Surrogate

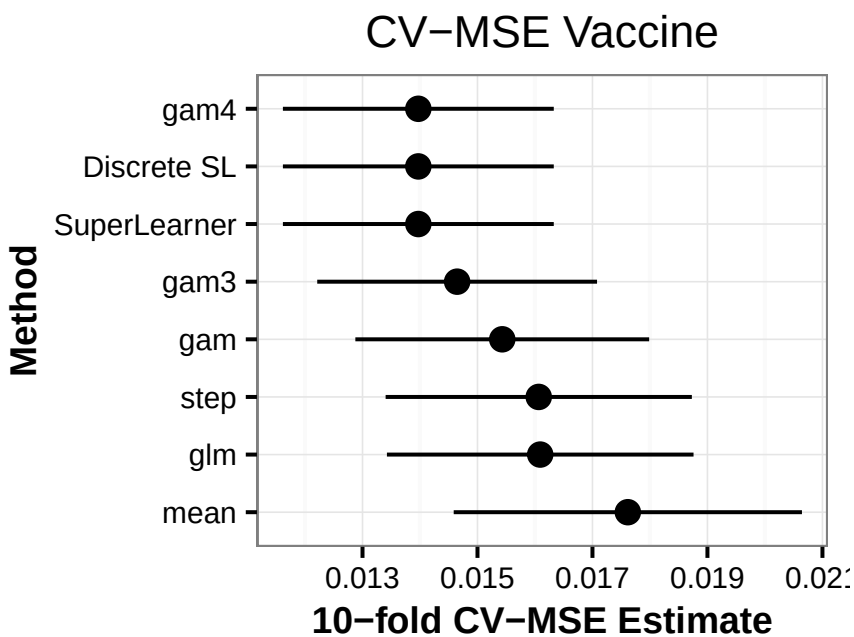
Table: Learning Algorithms Employed

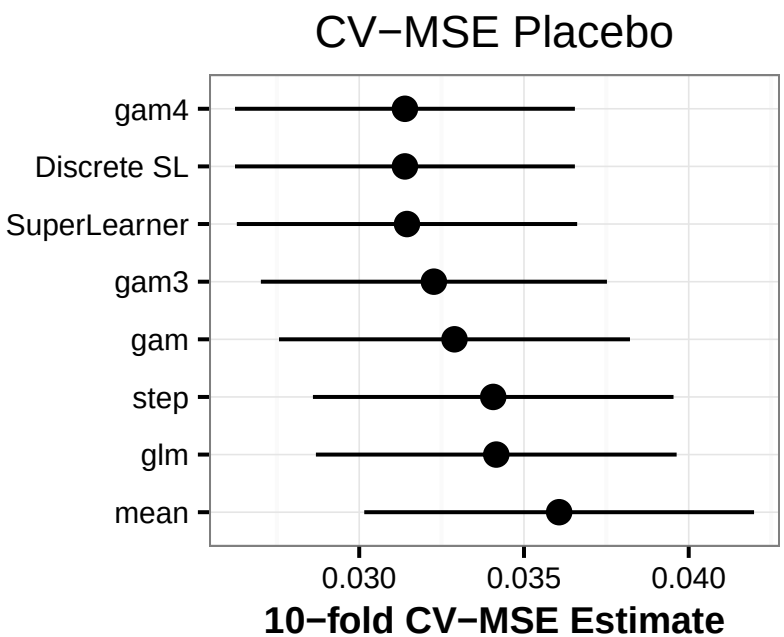
Learners	
mean:	$E(Y\|Z=z,W,S)=\beta_z$ for $z \in \{0,1\}$
LR:	Logistic regression with all input variables
step LR:	Best LR model by AIC through a step-wise search
gam2:	generalized additive model[a] with 2 degrees of freedom
gam3:	generalized additive model with 3 degrees of freedom
gam4:	generalized additive model with 4 degrees of freedom
discrete SL[b]	
super-learner[b]	

[a] Hastie and Tibshirani (1990) textbook

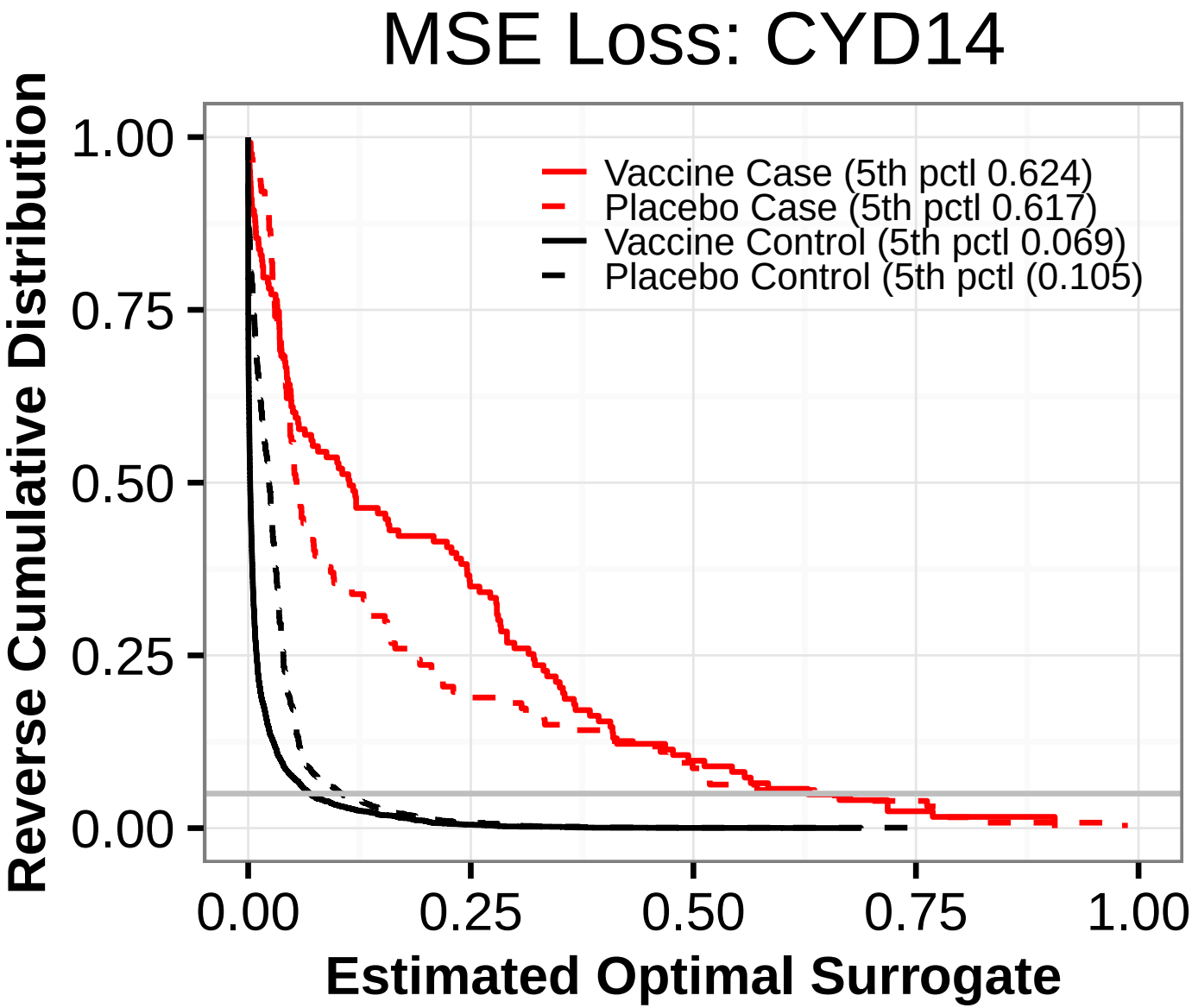
[b] van der Laan, Polley, and Hubbard (2007); van der Laan and Rose (2011) textbook

Cross-validated Mean-Squared Errors (CV-MSEs): CYD14





Empirical RCDFs for the Estimated Optimal Surrogate Values: CYD14

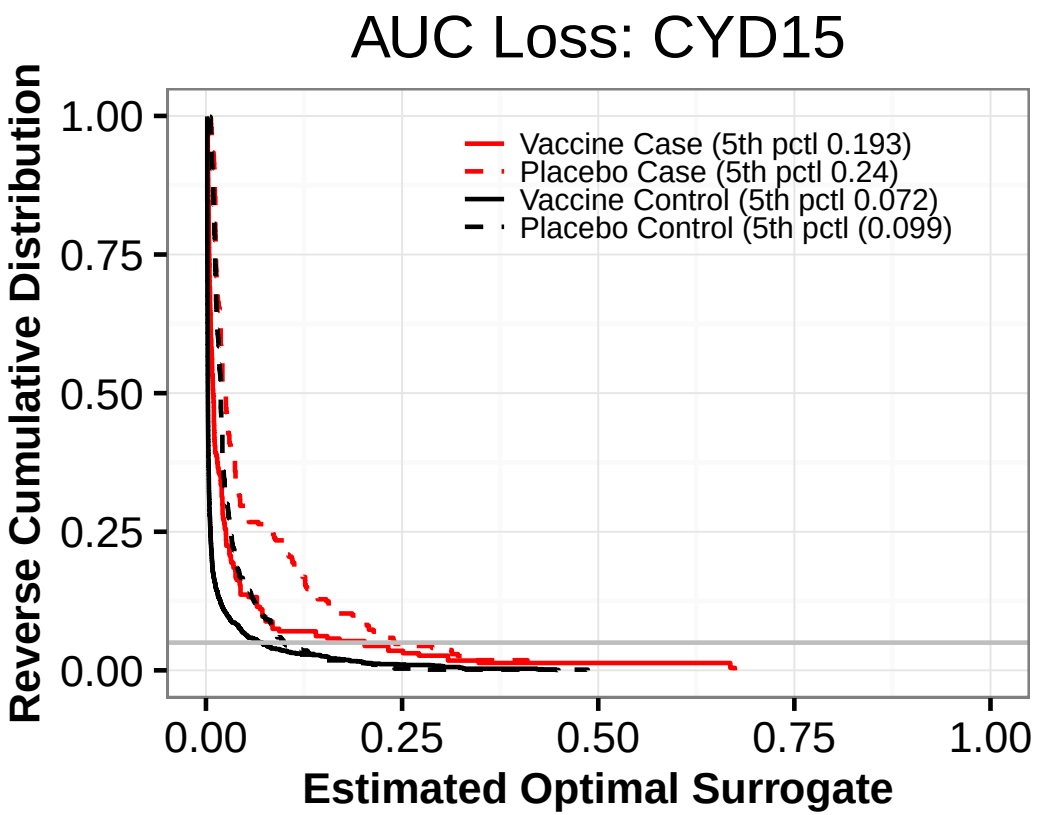


Estimated Optimal Surrogate TMLEs of $E[Y|Z=1]$, $E[Y|Z=0]$, and VE: CYD14

Parameter	TMLE	MLE using Y
$E[Y\|Z=1]$	1.8% (95% CI 1.5–2.1)	1.7% (95% CI 1.4–2.1)
$E[Y\|Z=0]$	3.7% (95% CI 3.1–4.4)	3.7% (95% CI 3.1–4.4)
$VE = 1 - \frac{E[Y\|Z=1]}{E[Y\|Z=0]}$	52% (95% CI 41–66)	55% (95% CI 43–68)

Using the Estimated Optimal Surrogate in CYD15

Calculate the estimated optimal surrogate endpoint $\psi_n^{TMLE}(W^*, Z^*, S^*)$ (built in CYD14) for all CYD15 participants– How well does it predict Y^*?



- Reduced classification accuracy for the new setting

How Well Does the Surrogate Estimate VE^* in CYD15?

Table: Estimation in CYD15 based on the estimated optimal surrogate $\psi_n^{TMLE}(W^*, Z^*, S^*)$ built in CYD14 (not using outcome data Y^* in CYD15) vs. estimation using Y^* in CYD15

Parameter	ψ_n^{TMLE}	MLE using Y^*
$E[Y^*\|Z=1]$	1.5% (95% CI 1.4-1.6)	1.8% (95% CI 1.4–1.9)
$E[Y^*\|Z=0]$	3.3% (95% CI 3.2–3.4)	3.7% (95% CI 3.6–4.5)
$VE^* = 1 - \frac{E[Y^*\|Z=1]}{E[Y^*\|Z=0]}$	54% (95% CI 44–67)	59% (95% CI 51–65)

Compare Predictive Ability of Input Variable Sets

Table: Cross Validated AUCs* with 95% CIs

Input Set	CYD14 Vaccine	CYD14 Placebo	CYD15 Vaccine	CYD15 Placebo
(1) Demographics	0.61 (0.57, 0.66)	0.6 (0.55, 0.65)	0.54 (0.5, 0.58)	0.5 (0.47, 0.54)
(2) All baseline	0.89 (0.86, 0.92)	0.79 (0.76, 0.83)	0.58 (0.54, 0.61)	0.55 (0.51, 0.58)
(3) Month 13 titers	0.71 (0.67, 0.75)	0.63 (0.58, 0.68)	0.65 (0.62, 0.69)	0.57 (0.54, 0.61)
(4) All data	0.89 (0.86, 0.91)	0.76 (0.72, 0.8)	0.78 (0.76, 0.8)	0.6 (0.57, 0.64)

*Cross-valided area under the ROC-curves (Van der Laan, Hubbard, and Pajouh, 2013)

- The user can judge the tradeoff of **accuracy** and **simplicity** of the estimated optimal surrogate

Checking Assumptions of the Transportability Theorem

Transportability Assumptions

1. A^* is randomized conditional on W^*
2. $E[Y^*|W^* = w, A^* = a, S^* = s] = E[Y|W = w, A = a, S = s]$ for all (w,a,s) in a support of (W^*, A^*, S^*)
3. A support of (W^*, A^*, S^*) is contained in a support of (W, A, S)

- *Condition 1* is met by the design of CYD15: both CYD14 and CYD15 randomized treatment
- *Condition 2* could be examined by comparing estimates of $E[Y^*|W^* = w, A^* = a, S^* = s] = E[Y|W = w, A = a, S = s]$
- *Condition 3*
  - CYD14 age range 2–14; CYD15 9–16 (assumption fails)
  - All titer variables had the same minimum values
  - Maximum titers also similar except Month 13 serotype 3 maximum titers 14% higher for CYD15 and baseline serotype 1 (4) maximum titers 18% (2%) greater for CYD15