Outline Talk 6

1. Introduction: Concepts and definitions of sieve effects / sieve analysis
   - Vaccine efficacy versus particular pathogen strains
   - Sieve effects and other effects
   - Some immunological considerations
   - Some sieve analysis results from HIV-1 vaccine efficacy trials

2. Some statistical approaches to sieve analysis
   - Binary endpoint (Infected yes/no)
     - Discrete pathogen types: Categorical data analysis
     - Continuous types: Distance-to-insert comparisons

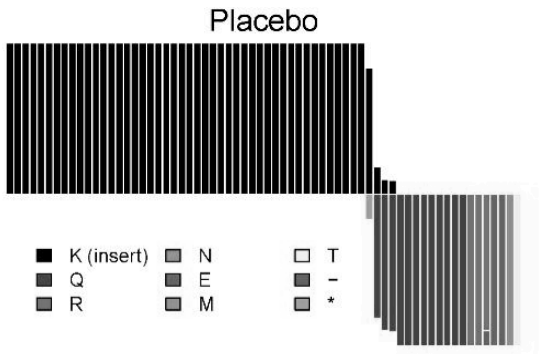
3. Assumptions required for interpretation as per-exposure vaccine efficacy

Overview of statistical approaches to sieve analysis

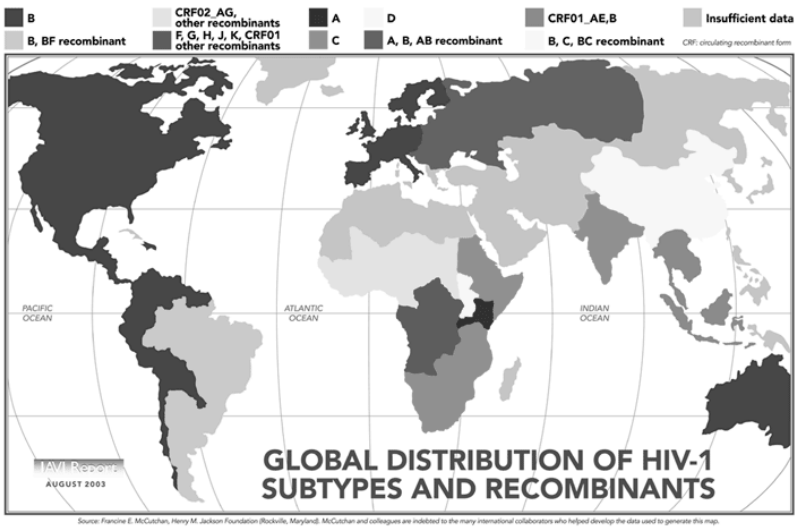
- Binary endpoint (Infected yes/no)
  - Discrete pathogen types: Categorical data analysis
  - Continuous types: Distance-to-insert comparisons
- Time-to-event endpoint (Survival analysis)
  - Discrete types: Competing risks
  - Continuous types: Mark-specific vaccine efficacy

Categorical Sieve Analysis

Fisher p = 0.3575	K	Q	R	N	E	M	T
Placebo	48	11	3	0	2	1	1
Vaccine	28	10	1	3	1	1	0

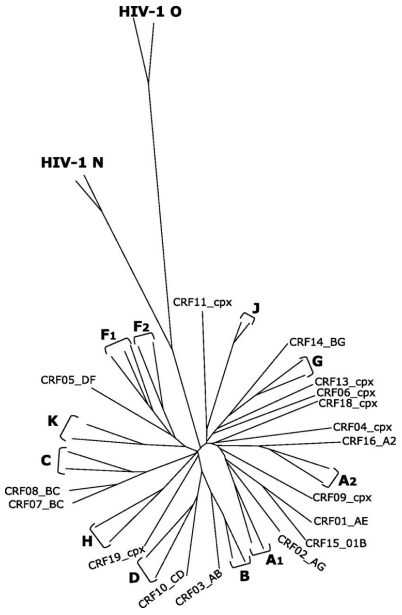


2003 Global Map of HIV-1 Subtypes



A vaccine might have different efficacy against different clades (subtypes) of HIV-1

Evolutionary relationships among nonrecombinant HIV-1 strains.



A vaccine might have different efficacy against different clades (subtypes) of HIV-1

Buonaguro L et al. J. Virol. 2007;81:10209-10219

Categories of pathogen types

- Human trials of preventative vaccines against heterogenous pathogens

Pathogen	Citation
Hepatitis	Szmuness et al. 1981
Cholera	Clemens et al. 1991 van Loon et al. 1993
Rotavirus	Lanata et al. 1989 Ward et al. 1992 Ukae et al. 1994 Jin et al. 1996 Rennels et al. 1996
Pneumococcus	Amman et al. 1977 Smit et al. 1977 John et al. 1984
Influenza	Govaert 1994
Malaria	Alonso et al. 1994

Some of these data are summarized in Gilbert et al. (2001, J Clin Epidem)

Vaccine efficacy vs pathogen type

- Human trials of preventative vaccines against heterogenous pathogens
  - Often there is no quantitative statistical assessment of differential VE across pathogen types
  - When there is, the interpretation and validity is often unclear
- Type-specific VE assessment
  - Can improve power to detect VE
  - Is often of interest
    - Multivalent vaccines: VE for each type
    - Partially protective vaccines: understanding and improving

Data setup

- Randomized vaccine trial
- K categories of infecting pathogens
  - (distinct strains, serotypes, amino acids, etc.)
  - Labeled 1 .. K
  - *wlog*, let category 1 be the "vaccine prototype strain"
    - eg the "insert" strain contained in the vaccine
- Nominal categorical: unordered strains
- Ordered categorical: eg ordered by distance to 1
- (Other methods consider continuous distances)

Meaningful classification

- Problem: sparsity of the 2xK table
  - in HIV, no clear serotypes
    - Star-like phylogeny within each clade
    - Each virus is unique (if you examine closely)
  - In general, for interpretation, want meaningful categories
- Solution: add structure to the table
  - Categorize infecting strains into nominal groups
    - Putatively related to strain-specific VE
    - eg: subtype/clade
    - eg: phenotype (in HIV, tropism: X4 vs R5)
  - (or) Order infecting strains
    - by putative correlate of strain-specifc VE
    - eg: order by measure of similarity to vaccine insert strain
      - Substitution matrix for nucleotide or amino acid sequence
  - Also possible: multidimensional features

Categorical data for sieve analysis

- Data: a 2xK table of counts

	1	2	3	4	5	…	K
Placebo						…	
Vaccine						…	

- Some analysis approaches
  - Fisher's exact test (or Fisher-Freeman-Halton for > 2x2)
  - Bayesian / Multinomial modeling
  - Recode as continuous values, use *eg* t-test
  - Multinomial logistic regression

Multinomial Logistic Regression (Cox, 1970; Anderson, 1972)

$$\Pr(Y = s|V) = \frac{\exp\{\alpha_s + \beta_s v\}}{1 + \sum_{k=2}^{K} \exp\{\alpha_k + \beta_k v\}}$$

- $s \in 1, \ldots, K$
- $\alpha_1 = \beta_1 \equiv 0$
- $v = 1$ indicates vaccine recipients

- A generalized linear logit model

$$\log\left\{\frac{\Pr(Y = s|v)}{\Pr(Y = 1|v)}\right\} = \alpha_s + \beta_s v$$

- Interpretation of the regression coefficients

$$\begin{aligned}\beta_s &= \log\left\{\frac{\Pr(Y = s|\text{vacc})}{\Pr(Y = 1|\text{vacc})} \Big/ \frac{\Pr(Y = s|\text{plac})}{\Pr(Y = 1|\text{plac})}\right\} \\ &= \log\{\,\text{OR}(s)\,\}\end{aligned}$$

Multinomial Logistic Regression model properties

- Minimal assumptions
- Estimation by maximum likelihood
- Exact methods an option
  - Hirji, K. F. (1992). Computing exact distributions for polytomous response data. *Journal of the American Statistical Association* **87**, 487-492.
- Easily extended to ordered categories
  - Anderson's (1984) ordered stereotype model
  - Same model, but use $\beta_s = \phi_s \beta$ and set $\phi_1 \equiv 0$
  - For monotonicity, constrain the order, *eg*

$$0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_K = 1$$

Multinomial Logistic Regression alternative ordered models

- Cumulative strain categories model
  - McCullagh 1980

$$\frac{\Pr(Y > s|v)}{\Pr(Y \leq s|v)} = \exp\{\alpha_s + \beta_s v\} \qquad s \in 1, \ldots, K-1$$

  - Interpretation of the regression coefficients

$$\exp\{\beta_s\} = \frac{\Pr(Y > s|\text{vacc})/\Pr(Y > s|\text{plac})}{\Pr(Y \leq s|\text{vacc})/\Pr(Y \leq s|\text{plac})}$$
$$= \text{OR}(> s)$$

- Scored ordinal models
  - Replace β_s with $(s-1)\beta$
  - Scored models have increased precision
    - But stronger modeling assumptions

Nonparametric Tests for Differential VE

- Null hypothesis: all $\text{OR}(s) = 1$
- Nominal categorical:
  - Likelihood ratio chi-squared test (Armitage 1971)
- Ordinal categorical:
  - Test for trend in strain-specific odds ratios
  - Breslow and Day (1980)
- Multiple vaccine dose groups:
  - Kruskal-Wallis test
  - Linear-by-linear association test (Agresti, 1990, p. 284)

Parametric Tests for Differential VE

- MLR or cumulative categories
  - Null hypothesis: all $\beta_s = 0$
  - Likelihood ratio chi-squared test
  - Zelen's test (1991)
  - Note: could also test null that a subset of the $\beta_s = 0$
- Categorical scored models
  - Null hypothesis: $\beta = 0$
- Continuous Model
  - Null hypothesis: $\beta = 0$
  - Likelihood ratio, Wald, and score test

Hepatitis B example

- Hepatitis B vaccine trial in New York
  - Szmuness et al., 1981
- MLR test of differential VE
  - Sieve LRT: $\chi^2_2 =$ **30.2** $p < 10^{-6}$
  - Zelen's: $\chi^2_2 = 26.1, p < 10^{-5}$
- MLR parameter estimates

	Hep B	Hep A	Hep other
Placebo	63	27	11
Vaccine	7	21	16

$$\exp(\hat{\beta}_2) = \frac{\text{RR(hep A)}}{\text{RR(hep B)}} = 7.0 \qquad 95\%\ \text{CI: } (2.7, 18.4)$$

$$\exp(\hat{\beta}_3) = \frac{\text{RR(hep other)}}{\text{RR(hep B)}} = 13.1 \qquad 95\%\ \text{CI: } (\boxed{\textbf{4.4, 39.1}})$$

HIV-1 Ordinal Categorical Example
The 'GPGRAF' V3 loop tip sequence

- VaxGen's MN/GNE8 gp120 vax; early-phase trial
  - See Gilbert, Self, Ashby 1998
  - Not randomized
  - Low power, few endpoints
    - Breslow-Day: p=0.11
    - Kruskal-Wallis: p = **0.13**

# mismatches	0	1	>1
Historical	43	20	4
Vaccine	2	1	2

Fit of sieve models to breakthroughs in Genentech vaccine trial

Model	Category	$\hat{\beta}$	SE($\hat{\beta}$)	$\exp\{\hat{\beta}\} = \widehat{OR}$	95% CI[a] $\widehat{OR}$	p-value
MLR	1	.072	1.25	1.07	(0.09, 12.56)	0.95
	2	2.38	1.13	10.75	(1.18, 98.16)	0.035
Cumulative logit	>0	0.99	0.95	2.69	(0.42, 17.22)	0.30
	>1	2.35	1.05	10.50	(1.35, 81.96)	0.025
Adjacent categories	1	1.12	0.63	3.06	(0.90, 10.43)	0.074
linear logit	2	2.24	1.26	9.35	(0.80, 108.69)	0.074
Proportional odds	>0	1.18	0.52	3.27	(1.17, 9.11)	0.024
	>1	1.18	0.52	3.27	(1.17, 9.11)	0.024

[a] Ninety-five percent confidence intervals are derived from a normality approximation and the observed inverse information matrix.

Generalized Logistic Regression Model
(Gilbert et al, 1999; Gilbert, 2000)

- Continuous analog of the MLR model
  - Parameterized $\beta_s = g(s)\theta \qquad s \in [0, \inf)$
    - For some deterministic function g

$$\Pr(Y = y|\text{vacc}) = \frac{\exp\{g(y)\theta\}f(y)}{\int_0^\infty \exp\{g(z)\theta\}dF(z)}$$

    - Where $f(y) \equiv \Pr(Y = y|\text{plac})$
  - Parametric component: regression coefficients
  - Nonparametric component:

    the placebo-recipient distribution F

Generalized Logistic Regression Model (continued)

- Interpretation of the regression coefficients

$$g(y)\theta = \log\{\mathrm{OR}(y)\} = \log\left\{\frac{\mathrm{RR}(y)}{\mathrm{RR}(0)}\right\}$$

  – Can also compute arbitrary log-odds ratios via:

$$\big(g(y_1) - g(y_2)\big)\theta = \log\left\{\frac{\mathrm{RR}(y_1)}{\mathrm{RR}(y_2)}\right\}$$

  – eg if $g(y) = y$,

$$\mathrm{RR}(y+1) = \exp\{\theta\}\mathrm{RR}(y)$$

Multidimensional pathogen variation

- The MLR and GLR models can accommodate pathogen variation described by multiple features
- Data examples:
  – Cholera: biotype, serotype, disease severity
  – Rotavirus: serotype, disease severity
  – HIV-1: vast possibilities
    - tropism
    - sequence distances to multiple vaccine inserts
    - presence (or affinity) of antibody binding targets
    - sequence distances in multiple regions

Multivariate GLR Model

- $\mathbf{Y} = (Y_1, \ldots, Y_d) \in [0, \infty)^d$
- *eg* for $d = 2$:

$$\Pr(\mathbf{Y} = (y_1, y_2) | \text{vacc}) = \frac{\exp\{g_1(y_1)\theta_1 + g_2(y_2)\theta_2 + g_1(y_1)g_2(y_2)\theta_3\} f(y)}{\int_0^\infty \int_0^\infty \exp\{g_1(z_1)\theta_1 + g_2(z_2)\theta_2 + g_1(z_1)g_2(z_2)\theta_3\} dF(z_1, z_2)}$$

- Can investigate dependence of VE on marginal distances, adjusting for other distances
  - *eg* $\frac{\text{RR}(y_1)}{\text{RR}(y_1')}$ adjusted for Y_2
- Can investigate interactions, *eg* does

$$\text{VE}(Y_1, Y_2) = \text{VE}(Y_1)\text{VE}(Y_2)?$$

HIV-1 Merck adenovirus-5 vector example

- Includes HIV-1 proteins coded by genes
  - *gag*, *pol*, and *nef*
- $\mathbf{Y} = (Y_{gag}, Y_{pol}, Y_{nef})$
  - Y_{gag} : a distance metric based on the *gag* gene
  - Y_{pol} : a distance metric based on the *pol* gene
  - Y_{nef} : a distance metric based on the *nef* gene
- Investigate how vaccine efficacy depends on heterogeneity in *gag*, *pol*, and *nef*

HIV-1 Merck adenovirus-5 vector example continued: introducing CDX metrics

- Question: What are the roles of CD4+ and CD8+ T-cell immune responses in vaccine protection?
  - Helper t cells (CD4+) vs Killer t cells (CD8+)
- $\mathbf{Y} = (Y_{\text{CD4+}}, Y_{\text{CD8+}})$ are **phenotypic marks**
  - $Y_{\text{CD4+}}$: strength of the CD4+ T cell response
    - **a T help metric**
  - $Y_{\text{CD8+}}$: strength of the CD8+ T cell response
    - **a CTL metric**
- Putting these together, get 3x2 = 6 dimensions:

$$\mathbf{Y} = (Y_{\text{CD4+},gag}, Y_{\text{CD4+},pol}, Y_{\text{CD4+},nef}, Y_{\text{CD8+},gag}, Y_{\text{CD8+},pol}, Y_{\text{CD8+},nef})$$

The *s*-sample GLR model

- *s* distinct covariate groups $x_1, \ldots, x_s$
  - *eg* for placebo & vaccine groups, $g_{\text{plac}}(y) \equiv 0$ in:

$$\Pr(Y = y|x_i) = \frac{\exp\{g_i(y)\theta\} f(y)}{\int_0^\infty \exp\{g_i(z)\theta\} dF(z)}$$

- For the *d*-dimensional case, the *s*-sample GLR model is

$$\Pr(Y = y|x_i) = \frac{\exp\{\sum_{k=1}^d g_{ik}(y)\theta_k\} f(y)}{\int_0^\infty \exp\{\sum_{k=1}^d g_{ik}(z)\theta_k\} dF(z)}$$

- *s* could also be multiple vaccine dose levels, stratification variable levels, etc.

Estimation for the GLR model

- The *s*-sample GLR model is a special case of a *semiparametric biased sampling model*:

$$\Pr(Y = y|i) = \frac{w_i(y,\theta)f(y)}{\int_0^\infty w_i(z,\theta)dF(z)} \qquad i \in 1,\ldots,s$$

- *eg* two-sample one-dimensional GLR model:

$$w_1(y,\theta) \equiv 1 \text{ and } w_2(y,\theta) \equiv g(y)\theta$$

- MLEs are obtained by maximizing a partial likelihood
  - see Gilbert et al, 1999 and Gilbert, 2000

Properties of the MLE in the GLR model

- GLR model is identifiable
- GLR model is uniquely estimable
  - Log profile partial likelihood is strictly concave
- MLEs are uniformly consistent, asymptotically Normal, asymptotically efficient
- Confidence intervals and variance estimation
  1. sample estimator of generalized Fisher information
  2. bootstrap
- Satisfactory finite-sample properties
- **Comparable to MLE in Cox model**

Outline Talk 6

1. Introduction: Concepts and definitions of sieve effects / sieve analysis
   - Vaccine efficacy versus particular pathogen strains
   - Sieve effects and other effects
   - Some immunological considerations
   - Some sieve analysis results from HIV-1 vaccine efficacy trials

2. Some statistical approaches to sieve analysis
   - Binary endpoint (Infected yes/no)
     - Discrete pathogen types: Categorical data analysis
     - Continuous types: Distance-to-insert comparisons

3. Assumptions required for interpretation as per-exposure vaccine efficacy

Model parameters and odds ratios

- Recall: for two-sample, one-dimensional MLR,

$$e^{\beta_s} = \frac{P_{vs}}{P_{v1}} \Big/ \frac{P_{ps}}{P_{p1}} = \frac{\mathrm{RR}(s)}{\mathrm{RR}(1)} = \log\{\,\mathrm{OR}(s)\,\}$$

$$P_{zs} \equiv \Pr\big(\text{ infected by strain } s \,\big|\, \text{infected in } [0,\tau],\ \text{vaccine treatment assignment is } z\big)$$

  - MLR: $e^{\beta_2} = \mathrm{OR}(2), \cdots, e^{\beta_K} = \mathrm{OR}(K)$
  - Scored MLR: $e^{\beta} = \mathrm{OR}(2), \cdots, e^{(K-1)\beta} = \mathrm{OR}(K)$
  - Ordered stereotype:

$$e^{\phi_2\beta} = \mathrm{OR}(2), \cdots, e^{\phi_K\beta} = \mathrm{OR}(K)$$

  - Cumulative categories:

$$e^{\beta_2} = \mathrm{OR}(>1), \cdots, e^{\beta_K} = \mathrm{OR}(>K-1)$$

  - GLR: $e^{g(y)\beta} = \mathrm{OR}(y)$

Retrospective vs Prospective

- All of the methods (so far) condition on infection
  - A post-randomization subgroup
  - Potential for bias despite randomized design
- Gilbert, Self, Ashby (1998) define
  - (have) *retrospective* relative risk

$$\mathrm{RR}(s) = \frac{\Pr(\text{ infected by strain } s \mid \text{ infected, vaccine recipient })}{\Pr(\text{ infected by strain } s \mid \text{ infected, placebo recipient })}$$

  - (want) *per-contact* relative risk

$$\mathrm{RR}^{\mathrm{pc}}(s) = \frac{\Pr(\text{ infected by strain } s \mid \text{ one exposure to strain } s, \text{ vaccine recipient })}{\Pr(\text{ infected by strain } s \mid \text{ one exposure to strain } s, \text{ placebo recipient })}$$

The Sieve Conditions

- *per-contact* RR **is** *retrospective* RR **if**

  (during the trial follow-up period)

  1. Infection is possible from at most one strain
  2. The relative prevalence of strains is constant
  3. Exposure distributions are the same in both treatment groups, and homogeneous across subjects*
- Proof in Gilbert, Self, Ashby (1998)
  - Holds for all of the aforementioned models
  - * the homogeneity aspect of this assumption can be relaxed. See Gilbert, Statistics in Medicine 2000.
  - See Gilbert, et al (2001) for more discussion
- Allows for the interpretation of strain-specific VE as prospective, per-contact-by-*s* VE