A Review of Statistical Methods for Evaluating Etiologic Heterogeneity

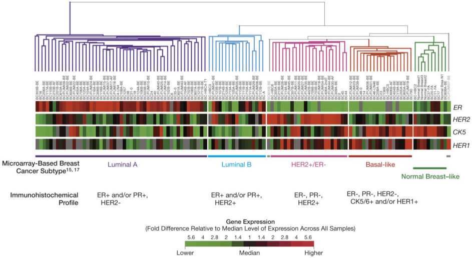
Emily C. Zabor & Colin B. Begg

Memorial Sloan Kettering Cancer Center

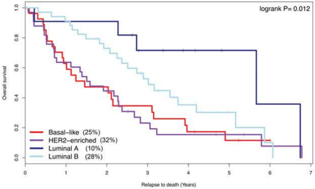
COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

May 12, 2016

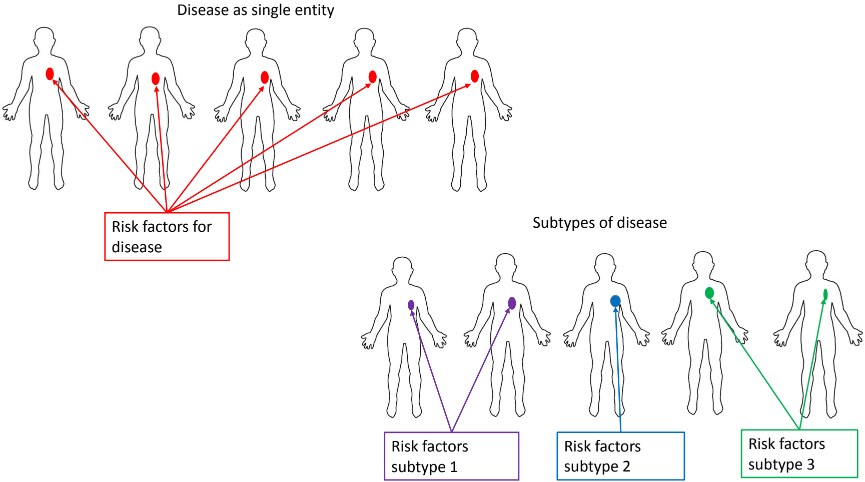
Molecular subtyping



→



Disease risk



Disease as single entity

Risk factors for disease

Subtypes of disease

Risk factors subtype 1

Risk factors subtype 2

Risk factors subtype 3

Included methods

Included methods for case-control studies:

1. Polytomous regression
2. Case-only polytomous regression
3. Two-stage extensions of polytomous regression
4. Methods that integrate subtyping with tests of heterogeneity

CASH data

- Cancer and Steroid Hormone Study (CASH)
- 551 cases, 2990 controls
- Collected data on $>$ 200 gene expression values and complete set of breast cancer risk factors
- Focus here on ER and PR for simplicity
- Cross-classify cases into 4 subtypes based on ER and PR status:

	PR-	PR+
ER-	201 (39%)	23 (4%)
ER+	51 (10%)	243 (47%)

CASH data

Typical question of interest: Is the effect of risk factors the same across all disease subtypes?

Here we focus on a single risk factor, parity, for simplicity.

Polytomous regression

Dubin and Pasternack, AJE 1986; 123(6):1101-17

$$\Pr(Y_i = m|X_i) = \frac{\exp(\alpha_m + \beta_m X_i)}{1 + \sum_{m=1}^{M} \exp(\alpha_m + \beta_m X_i)}, m = 1, \ldots, 4$$

- $exp(\beta_m) =$ odds ratio for parity as a risk factor for subtype m disease
- $H_0 : \beta_m = 0$ tests whether parity is associated with disease subtype m
- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ tests whether the association between parity and odds of cancer is the same across the four subtypes

Case-only polytomous regression

Begg and Zhang, CEBP 1994; 3(2):173-5

- All information needed to examine etiologic heterogeneity is contained in the cases
- Polytomous regression formula is the same as before
- Select one case group as the reference, for example ER-/PR-
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ tests whether parity is associated with disease subtype

Comparison of polytomous and case-only

	Odds ratios, $exp(\beta_m)$				
Method	ER-/PR-	ER+/PR-	ER-/PR+	ER+/PR+	p-value
Polytomous	0.66	1.77	1.96	1.30	0.030
Case only	ref	2.66	2.95	1.94	0.030

Comparison of polytomous and case-only

	Odds ratios, $exp(\beta_m)$				
Method	ER-/PR-	ER+/PR-	ER-/PR+	ER+/PR+	p-value
Polytomous	0.66	1.77	1.96	1.30	0.030
Case only	ref	2.66	2.95	1.94	0.030

Case-only ORs are simply ratios of polytomous ORs:

$$\frac{1.77}{0.66} = 2.66$$
$$\frac{1.96}{0.66} = 2.95$$
$$\frac{1.30}{0.66} = 1.94$$

Wang et al

AJE 2015; 182(3):263-270

- More general strategy to model multiple tumor factors and multiple risk factors
- First stage is a standard polytomous regression
- Second stage models the resulting regression parameters, $\hat{\beta}_m$ as:

$$\hat{\beta}_m = \gamma_0 + \sum_{k=1}^{K} \gamma_k w_{km} + e_m$$

- $H_0 : \gamma_k = 0$ tests whether the risk factor-subtype association changes over the levels of the kth tumor factor, holding all other tumor factors constant

Wang et al

AJE 2015; 182(3):263-270

$$\hat{\beta}_m = \gamma_0 + \sum_{k=1}^{K} \gamma_k w_{km} + e_m$$

γ_k is the ratio of OR for association between parity and subtype comparing levels of tumor factor k

m	$\hat{\beta}_m$	w_{1m}	w_{2m}	γs
1	$\hat{\beta}_1$	0	0	γ_0
2	$\hat{\beta}_2$	0	1	$\gamma_0 + \gamma_2$
3	$\hat{\beta}_3$	1	0	$\gamma_0 + \gamma_1$
4	$\hat{\beta}_4$	1	1	$\gamma_0 + \gamma_1 + \gamma_2$

Chatterjee

JASA 2004; 99(465):127-138

- Implements a two-stage approach to reduce parameter space
- A log-linear model is used at the second stage
- Tests whether the risk factor-subtype association differs by levels of the kth tumor factor when all other tumor factors are held constant
- Specialized estimation procedures handles missing risk factor data

Begg et al

Stat Med 2013; 32(29):5039-52

- Goal is to integrate identification of subtypes with measure of heterogeneity
- *k*-means clustering reduces the dimension of tumor factors
- Calculates a scalar measure of incremental explained risk variation, D, based on risk predictions from polytomous regression
- Can answer the question of how much heterogeneity across subtypes is explained by the set of risk factors as a whole

Yu et al

Biostatistics 2015; 16(1):5-16

- Binary recursive partitioning classifies patients into disease subtypes
- Considers K tumor factors but classification is only done based on a single risk factor
- Each split is selected to maximize heterogeneity with respect to the risk factor of interest

Conclusions

- Polytomous logistic regression is still the most widely applied and easily interpreted approach
- Traditional regression and two-stage approaches examine heterogeneity one risk factor and one tumor factor at a time
- Integrative approaches address reduction of tumor factor dimensionality
- Methods are needed that can handle both a large number of subtypes and a large number of risk factors