Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

Application of a method for identifying disease subtypes that are etiologically heterogeneous

Emily C. Zabor*,†, Shuang Wang†, and Colin B. Begg*

***Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center**
†Department of Biostatistics, Columbia University Mailman School of Public Health

Background

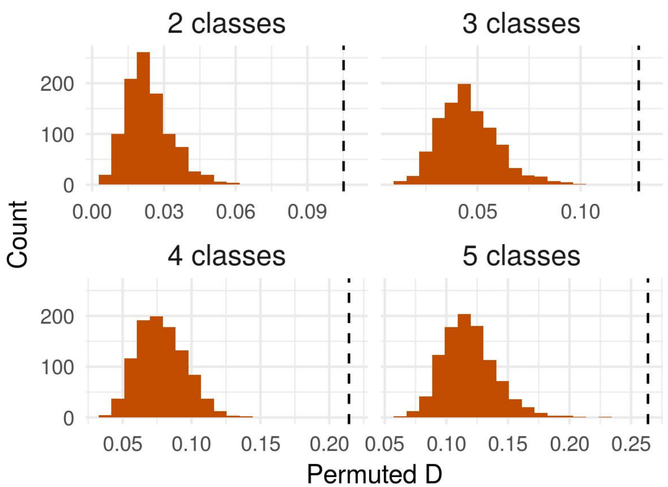
Given the known biological diversity of breast tumors, breast cancer could be considered a group of diseases with distinct etiologies and prognoses rather than a single disease. In previous work we proposed a novel clustering method to identify the most etiologically distinct subtypes based on high dimensional tumor marker data. Here we apply this approach to data from a large population-based breast cancer study.

Data

Data on 532 women with invasive breast cancer from Phase 3 of the Carolina Breast Cancer Study (CBCS) are used to discover the subtypes, based on a panel of 406 gene expression values. Subtypes are validated and risk factor effects examined in 482 cases and 1455 controls from CBCS Phases 1 and 2.

Results

For all class sizes, optimal D significantly exceeds the reference distribution. Based on 112/406 genes selected for inclusion with p<0.1 after correction for multiple testing.

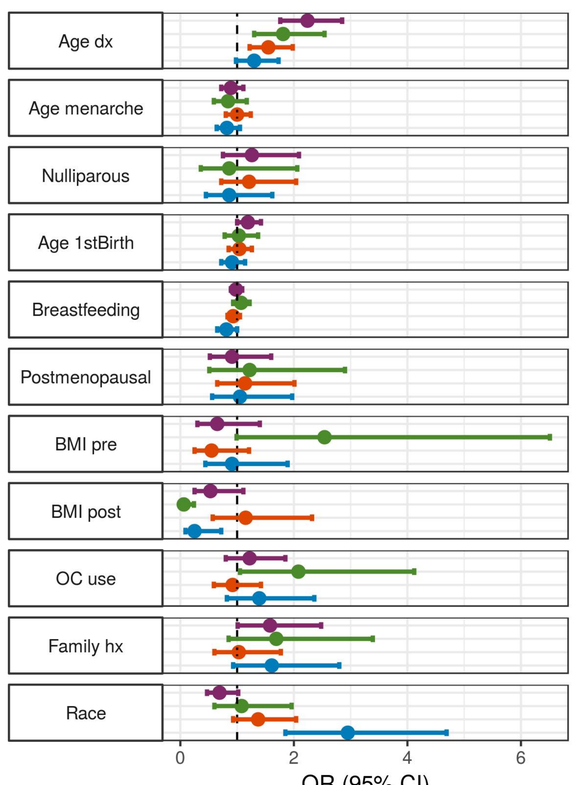


We select 4 as the optimal number of classes

Class size	D difference	P-value
3 vs 2	0.023	0.454
4 vs 2	0.109	0.002
5 vs 2	0.159	0.012
4 vs 3	0.086	0.007
5 vs 3	0.135	0.015
5 vs 4	0.049	0.358

Class size	D
Discovery	0.214
Validation	0.245
Traditional IHC 4-class	0.148

Age at diagnosis, menopausal status, and race differ across optimal subtype



Statistical Methods

1. Upfront selection of genes based on individual D and permutation-based p-values
2. *K*-means clustering of selected genes with 1000 random starts
3. Calculate D, a scalar measure of etiologic heterogeneity, for each unique solution
4. Select the class solution that maximizes D
5. Compare 2-class, 3-class, 4-class, and 5-class solutions with permutation tests
6. Estimate odds ratios and p-values with polytomous logistic regression for selected solution.

Conclusions

- The method can identify class solutions that demonstrate significant etiologic heterogeneity and validate reasonably
- Using D, we find a 4-class solution with greater etiologic heterogeneity than the traditional 4 classes of breast cancer
- ER is a key gene in distinguishing subtypes