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

Emily C. Zabor

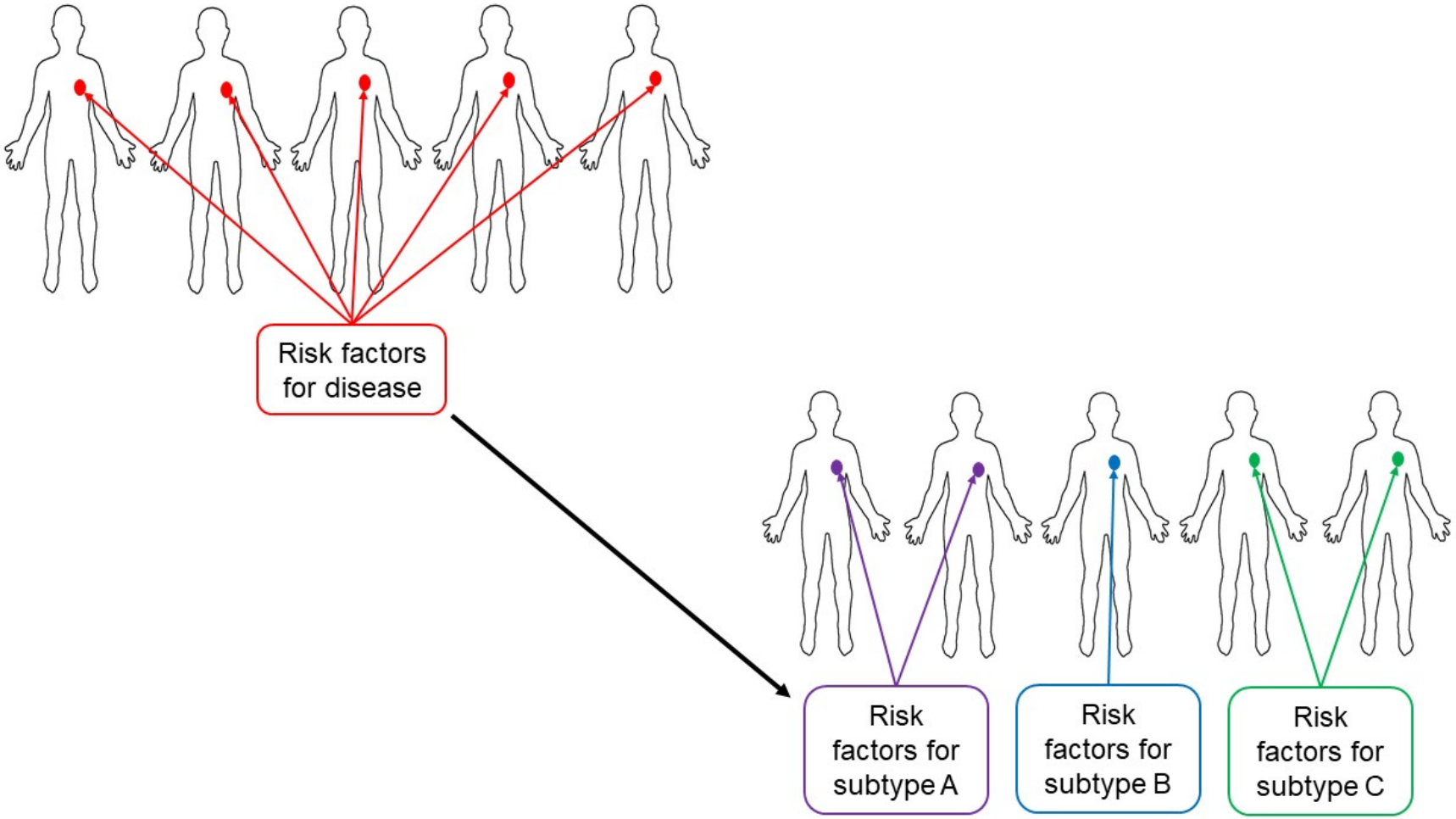
ENAR Meeting, March 27, 2018

Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

BIOSTATISTICS

Focus of epidemiologic research shifting from single disease by site to disease subtypes



Motivating data from the Cancer and Steroid Hormone (CASH) breast cancer case-control study

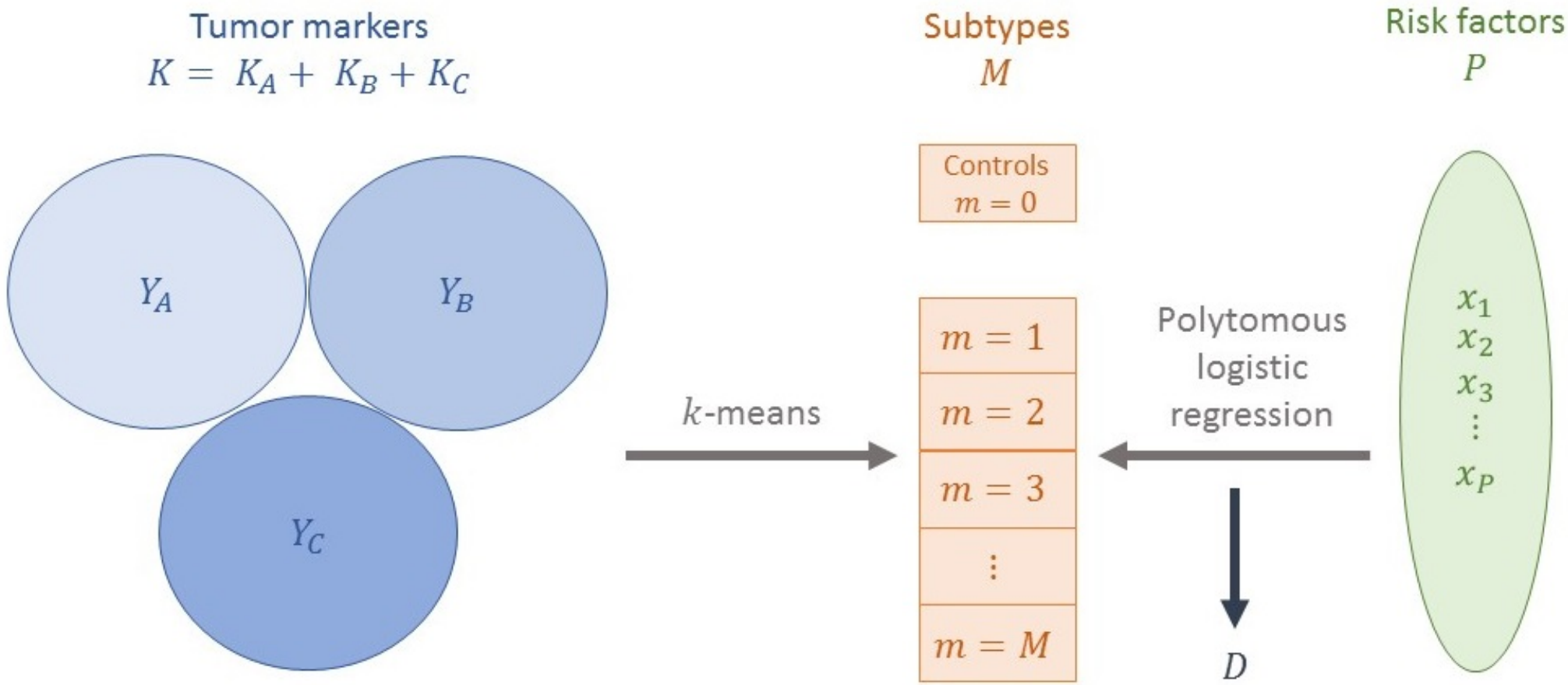
Data:

- 551 breast cancer cases
- 2990 population controls
- 12 risk factors for breast cancer
- 202 gene expression values (*cases only*)

Goal:

To discover the most etiologically heterogeneous disease subtypes with respect to relevant risk factors, according to high dimensional tumor marker data.

Proposed approach clusters marker data, optimizes etiologic heterogeneity



Risk heterogeneity is measured using the coefficient of variation and risk covariance

The incremental explained variation is defined as the difference between the mean explainable risk variation and the overall risk variation.

For three subtypes A, B, and C:

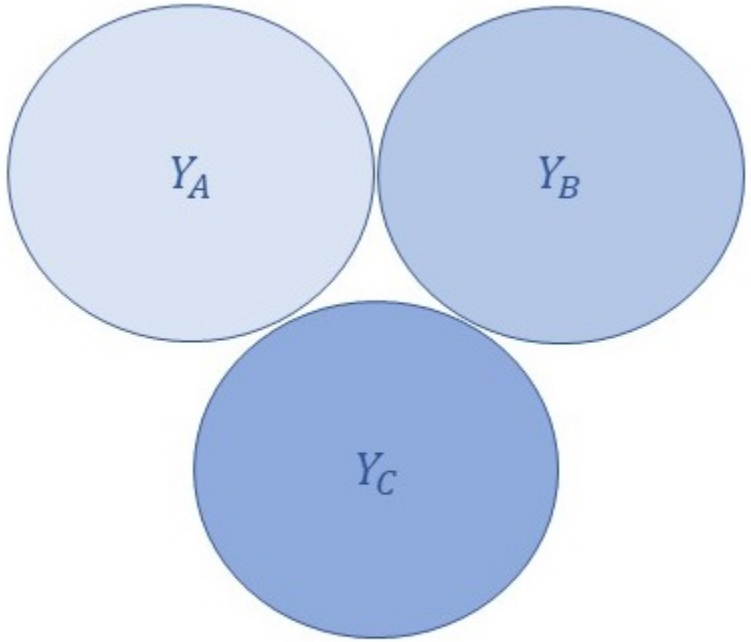
$$D = (\pi_A V_A^2 + \pi_B V_B^2 + \pi_C V_C^2) - V^2$$

where π_j is the relative frequency and V_j^2 is the coefficient of variation for subtype j.

But can this approach identify the subtypes that are truly the most etiologically heterogeneous?

Challenges:

1. Strength of structure of markers related to risk factors, Y_A



2. Presence of a subset of markers with structure unrelated to risk factors, Y_B

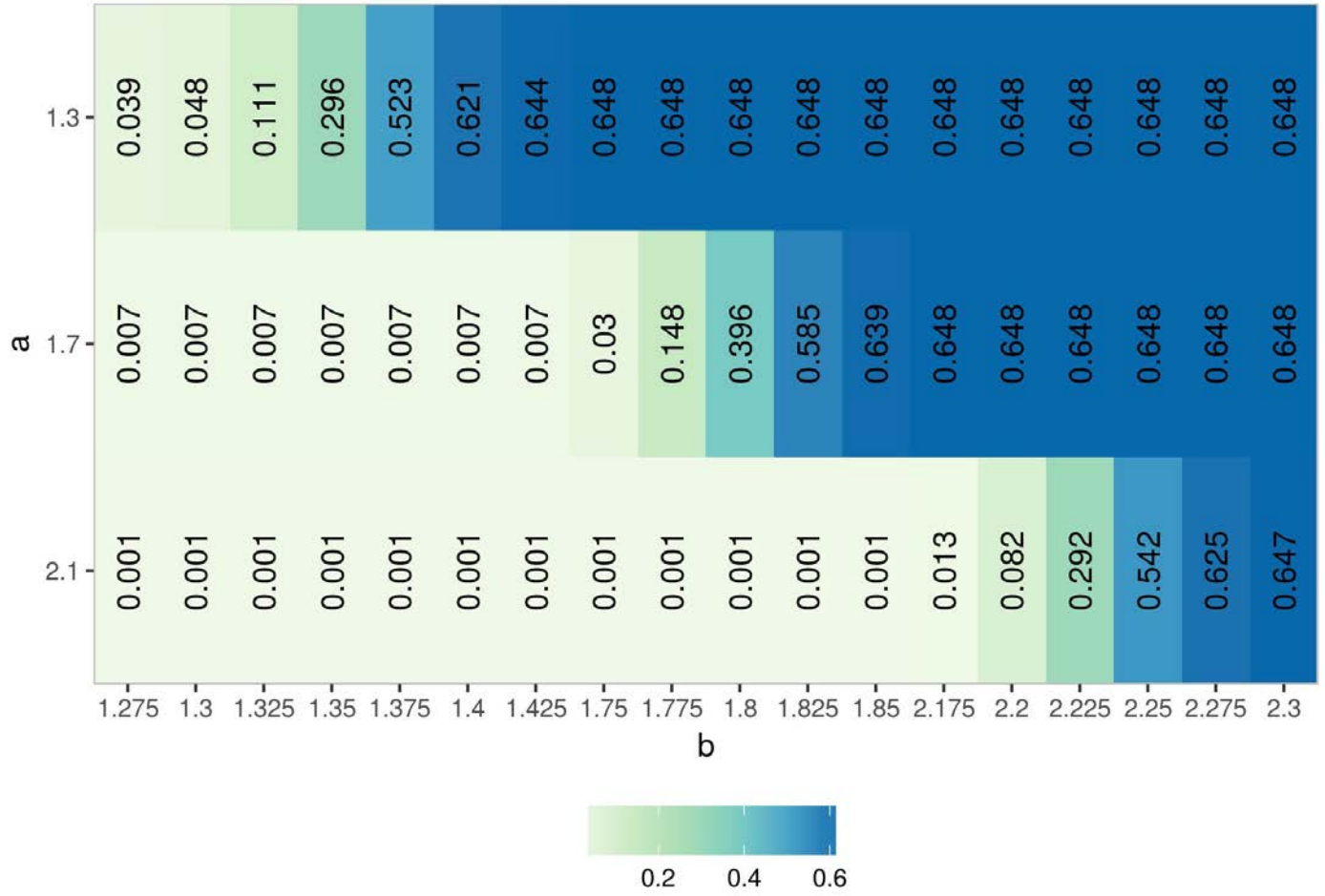
3. Presence of a large number of unstructured markers, Y_C

Simplified simulation settings allow for investigation of the statistical properties

- Risk factor data $X \sim N(\mu_m, I)$
- Selected tumor markers are correlated with specific risk factors, inducing etiologic heterogeneity
  - Related markers: $Y_A \sim N(\lambda_{Am}, I)$
  - Unrelated markers: $Y_B \sim N(\lambda_{Bj}, I)$
  - Unstructured markers: $Y_C \sim N(0, I)$

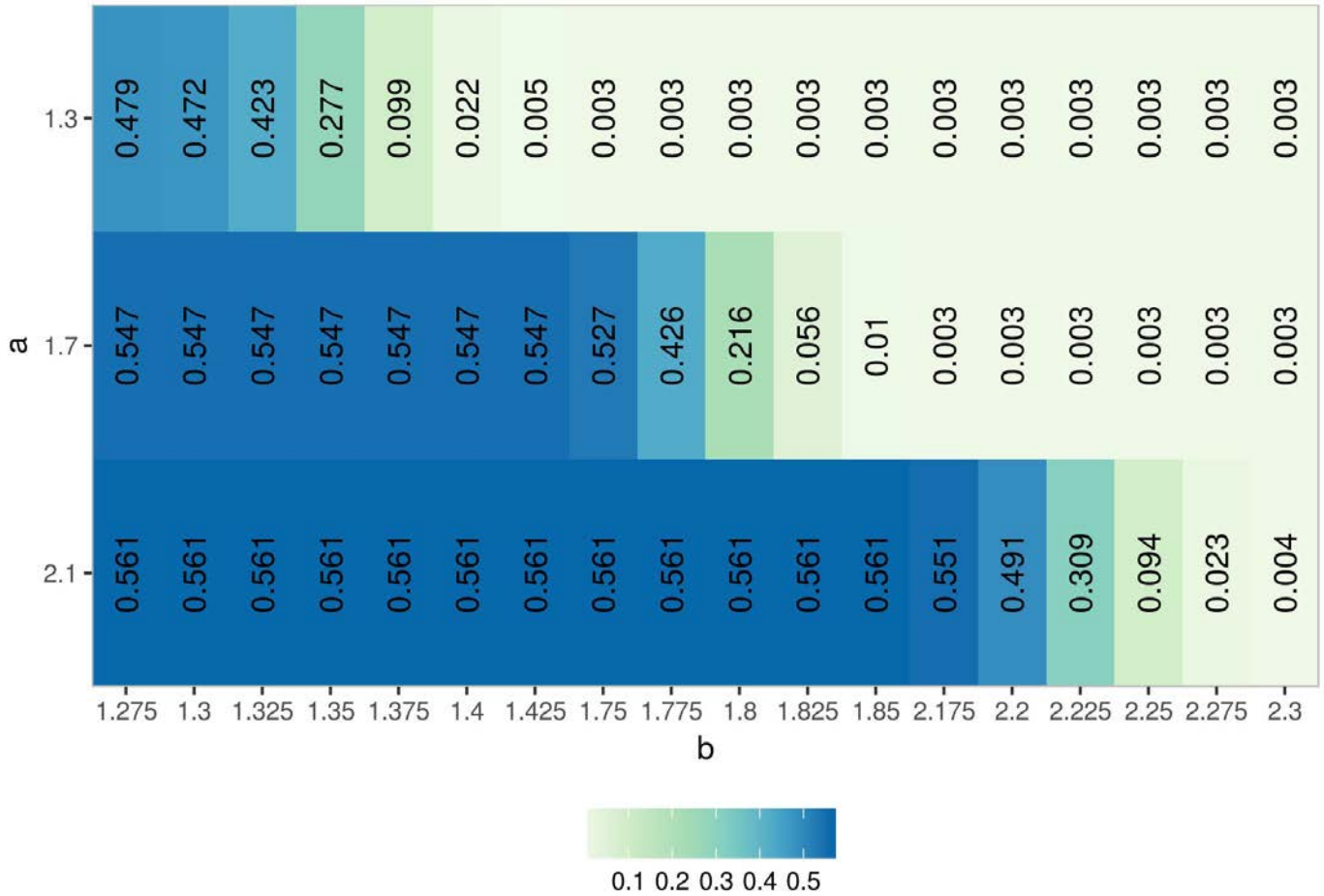
As strength in unrelated surpasses strength in related markers, misclassification increases

Clustering 15 related and 15 unrelated tumor markers



Average maximum D decreases as strength in unrelated surpasses strength in related markers

Clustering 15 related and 15 unrelated

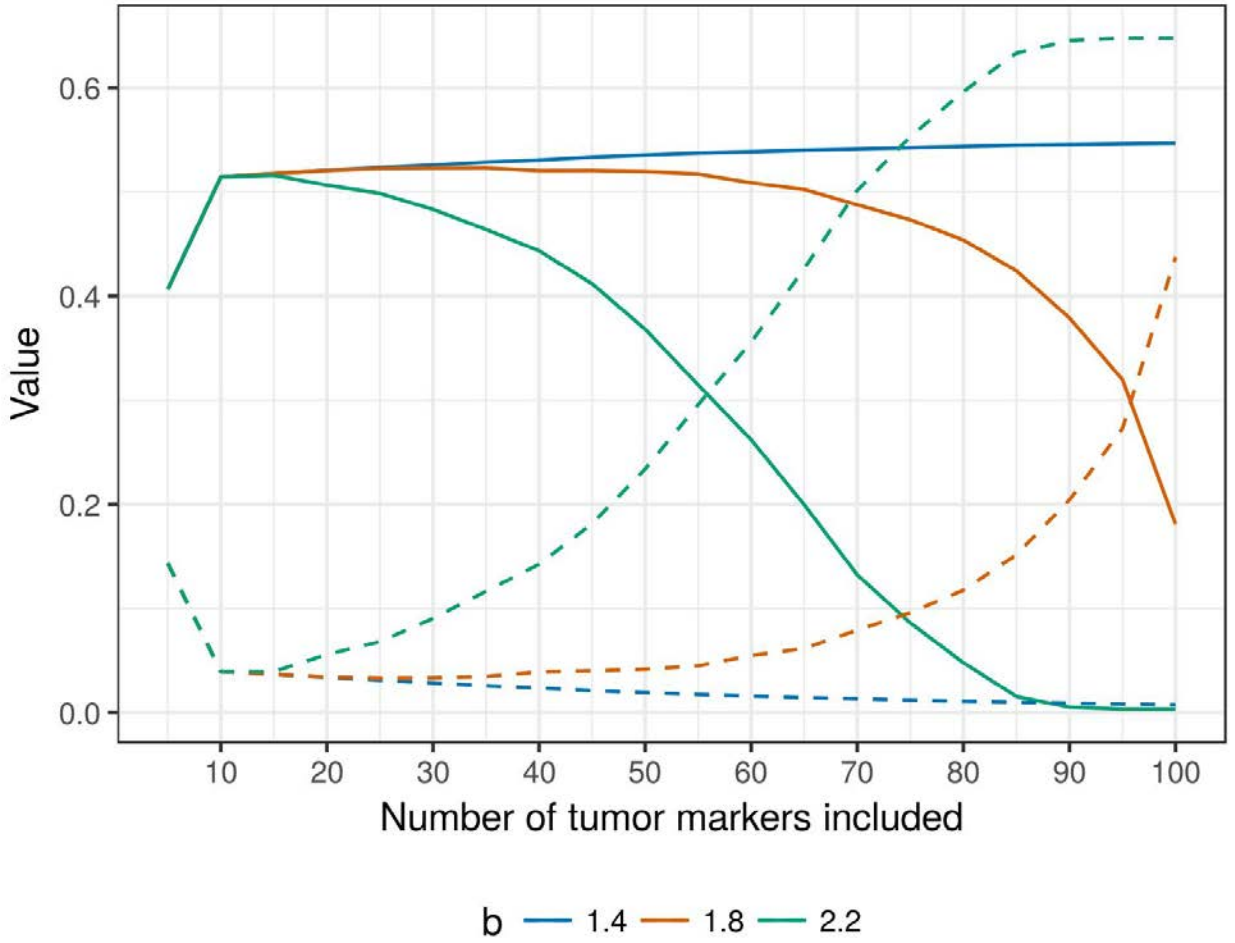


Variable selection is based on individual D for each marker

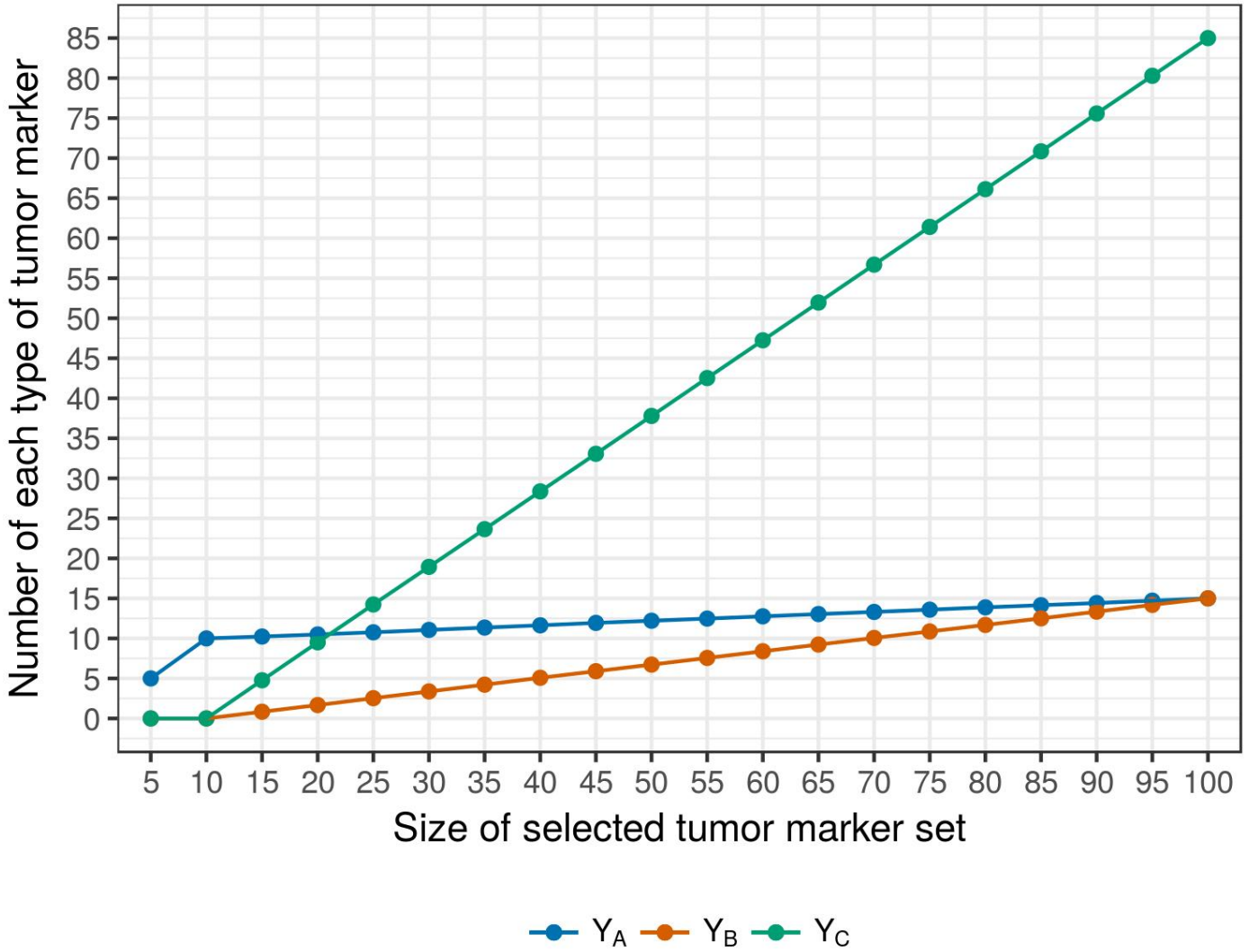
1. 15 related markers, 15 unrelated markers, and 70 noise markers
2. Individual D obtained for each marker by dichotomizing at median into two classes
3. Markers are ranked according to individual D
4. Set of included markers increased from 5 to 100, by 5

Misclassification increases and D decreases as the included number of tumor markers increases

Dashed lines denote misclassification, solid lines denote D

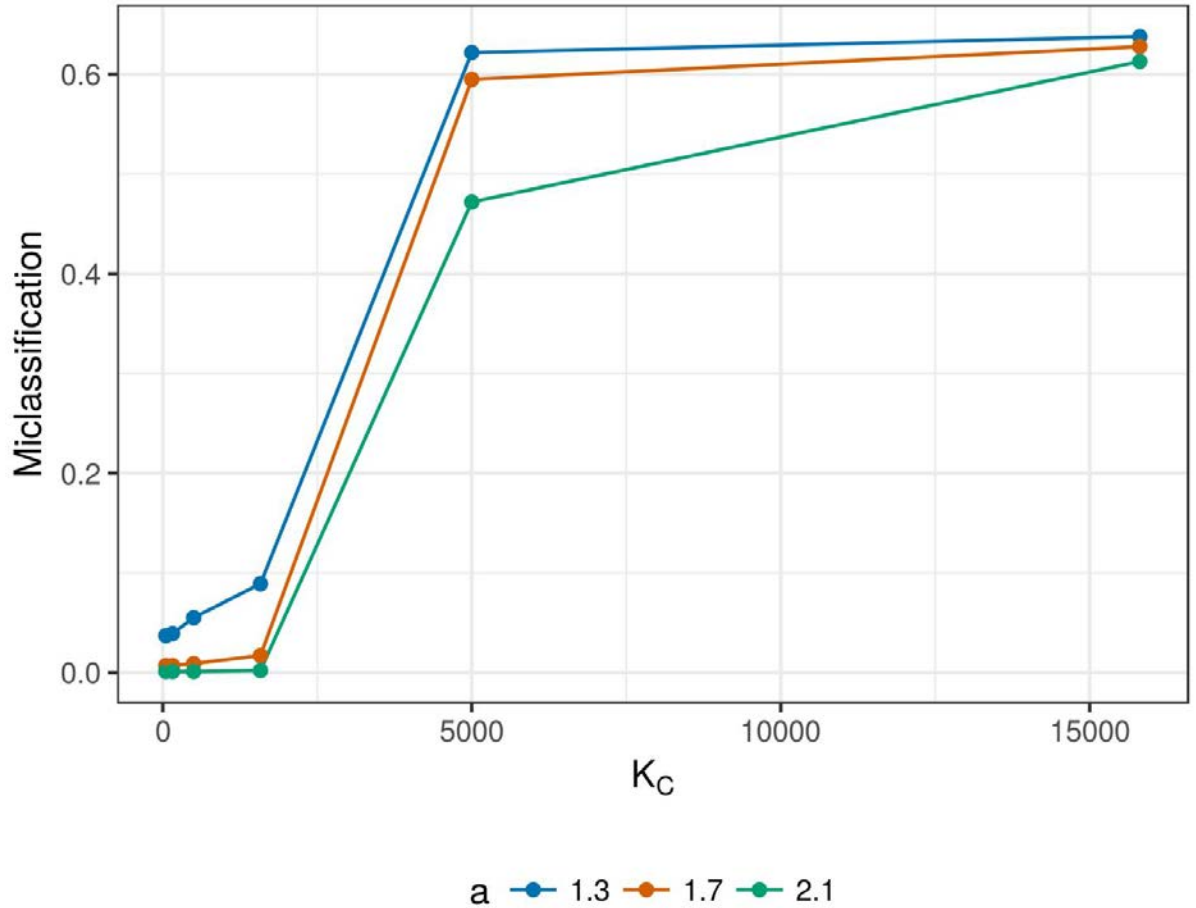


More noisy markers than markers with structure unrelated to risk factors tend to get selected



Increasing noisy markers eventually overwhelms even strong structure

Clustering 15 related, 15 unrelated, increasing number of noise markers



Pre-clustering variable selection improves alignment with traditional classes in CASH/CARE

33 genes selected based on adjusted permutation p-values

Full gene set	Her2-enriched	Luminal A	Luminal B	Triple neg
1	23	102	15	34
2	24	134	17	12
3	1	19	2	7
4	18	17	11	82
Reduced gene set	**Her2-enriched**	**Luminal A**	**Luminal B**	**Triple neg**
1	46	60	27	27
2	8	118	12	13
3	1	89	6	4
4	11	5	0	91

The method can find the true subtypes, and is enhanced by pre-clustering variable selection

- When structure in related markers is strong and structure in unrelated markers is relatively weak, we identify the true subtypes with low misclassification
- As the number of noisy markers increases, misclassification increases
- Pre-clustering variable selection improved performance across all settings
- Future work will explore determination of optimal number of subtypes

R package rickclustr available at https://github.com/zabore/riskclustr

Acknowledgements:

- Colin Begg, Memorial Sloan Kettering Cancer Center
- Shuang Wang, Columbia University
- Venkat Seshan, Memorial Sloan Kettering Cancer Center

Contact:

- Email: zabore@mskcc.org
- Slides: https://github.com/zabore/slidedecks

Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

BIOSTATISTICS