Multiple Chemical Exposures and Breast Cancer Risk: Findings from the California Teachers Study

Peggy Reynolds, David Nelson, Erika Garcia, Susan Hurley

Funding provided by NCI Grant R01 CA 77398 and DOD grant W81XWH-10-1-0134

University of California
San Francisco

HAPs and Breast Cancer Risk

- California Teachers Study Cohort
  - 112,378 participants in 1995
  - 5,676 breast cancers diagnosed 1995-2011
- US EPA National Air Toxics Assessment
  - Modeled hazardous air pollutants (HAPs)
  - Identified 24 mammary gland carcinogens (MGC)
- Analysis
  - HAPs defined by:
    - Individual MGC compounds
    - A summary measure by standardized concentrations
  - Cox proportional hazard models
  - Adjusting for breast cancer risk factors

Garcia et al., Environmental Health 14:14 2015

HAPs and Breast Cancer Risk

- Little evidence for risk associations for individual HAPs after adjustment for multiple testing

- Some evidence for risks in certain subgroups:
  - ER+/PR+ tumors: carbon tetrachloride
  - BMI <25: benzidine, carbon tetrachloride
  - Past/never HT use: carbon tetrachloride, ethylidene dichloride, vinyl chloride

- No evidence of risk for a summary measure of MGCs

Garcia et al., Environmental Health 14:14 2015

Approaches for assessing environmental chemical mixtures

We're not exposed to one chemical at a time but could consider:

- Mixtures by source
- Mixtures by chemical properties
- Mixtures by biologic effect
- *A priori* mixtures based on expert review
- Empirical testing

Empirical Testing: The challenge

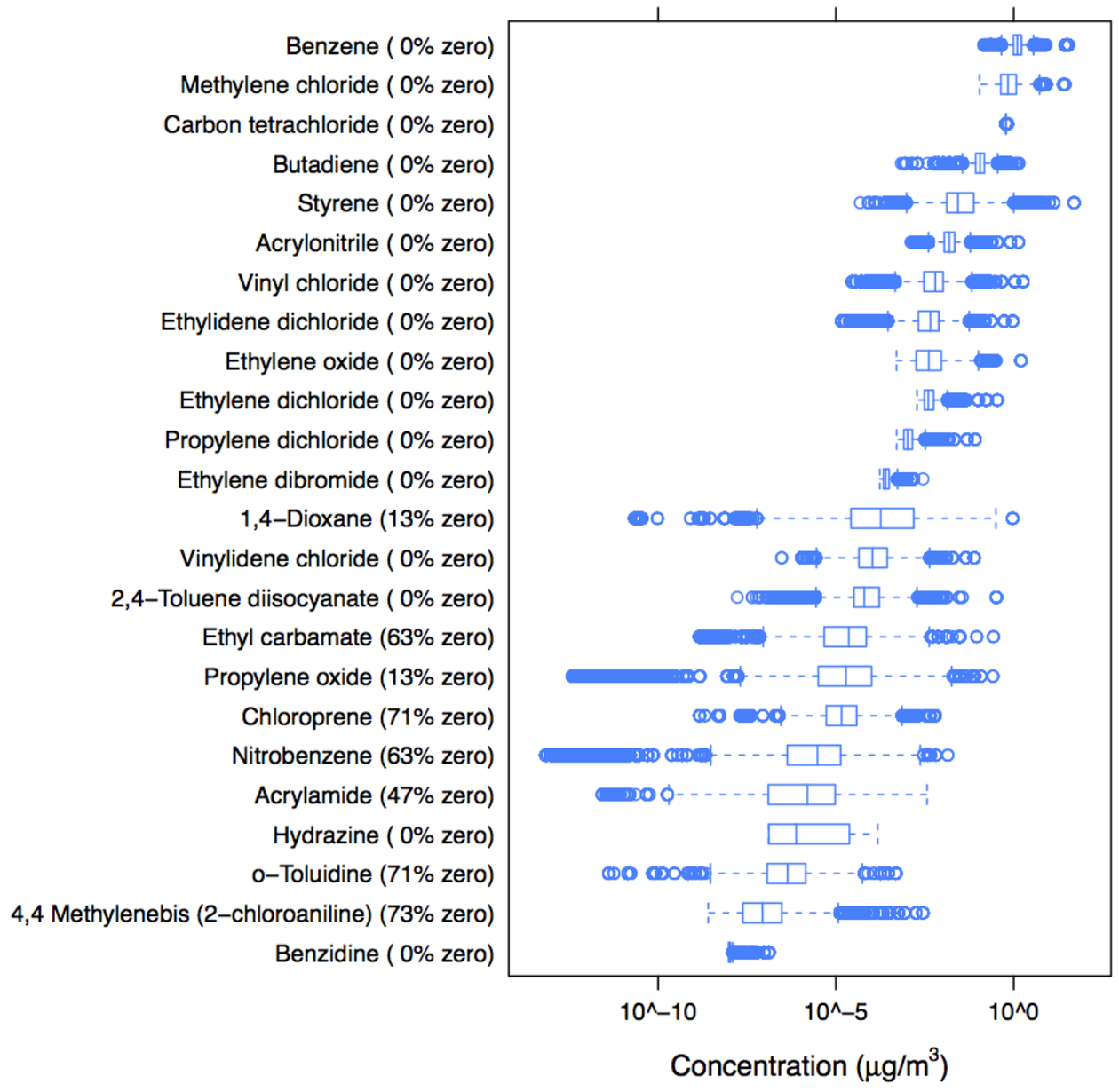
- What we want to know…
  - How exposure to hazardous air pollutants (HAPs) together contribute to breast cancer risk

- What we have…
  - EPA's National Air Toxics Assessment (NATA) census tract data from 2002
    - Modeled pollutant concentrations for >100 pollutants at the census tract level
  - Residence census tract at baseline for >100,000 California teachers in the California Teachers Study (CTS)

- A major challenge: how to account for pollutant correlations while estimating pollutant risks for BC?

Focus on 24 Mammary Gland Carcinogens (MGCs)

- The MGCs chosen based on prior expertise and literature search
- Over **10 orders of magnitude** difference in concentrations
  - use logs throughout
- Wide variation in percent zero:
  - Impute min/2 for zero to allow analyses based on logs
- Hence, need to standardize data for each MGC
  - Mean 0, variance 1



Two sources of correlation to account for:

- Spatial correlation between subjects
  - Observations for two subjects close to each other are likely to be more correlated that observations for two subjects far from each other
- Correlation between mixtures of similar pollutants
  - Concentrations for a given subject are likely to more correlated with each other

- Our focus on the *second* source of correlation
  - Although the first is just as likely to be present
  - Spatial analysis is a field all its own

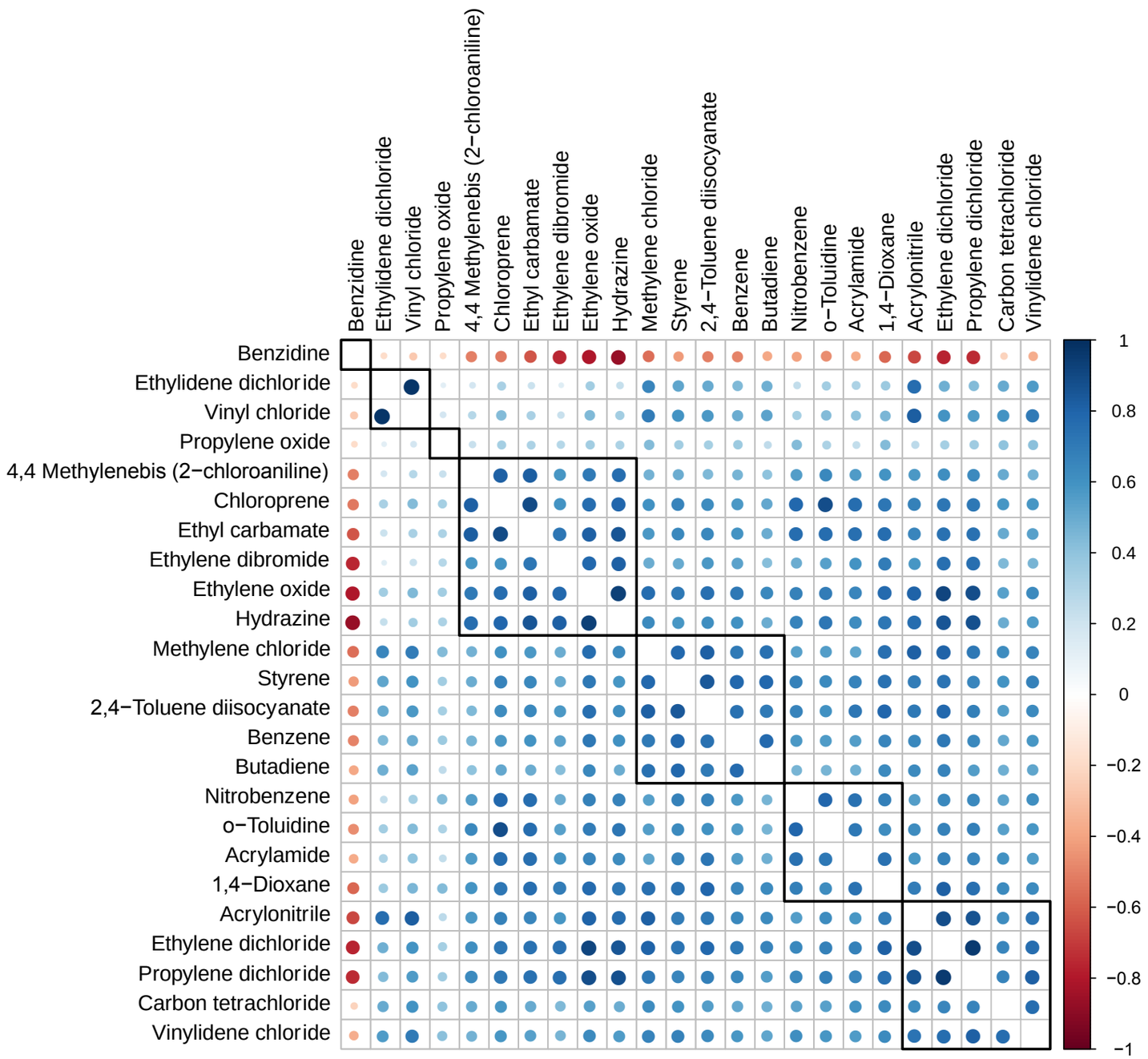
Problem: we had lots and lots of correlation

Correlation matrix of 24 MGCs based on Spearman rank correlations

- Benzidine negatively correlated with ***everything!***
- Groups of MGCs with high within-group correlation and lower between-group correlation

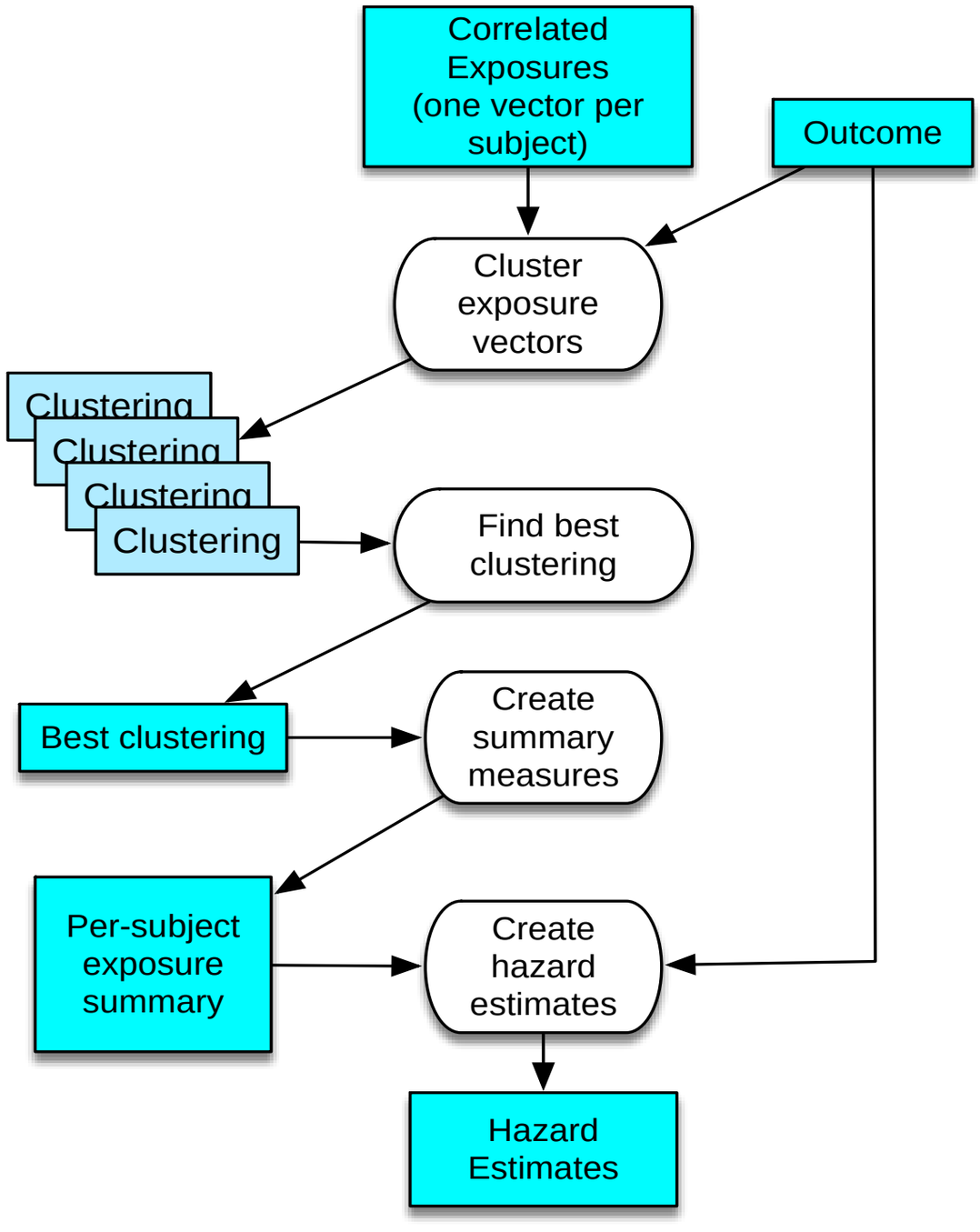
How do we summarize exposure-risk relationships in presence of such correlation?

- Are there subgroups that are interesting?



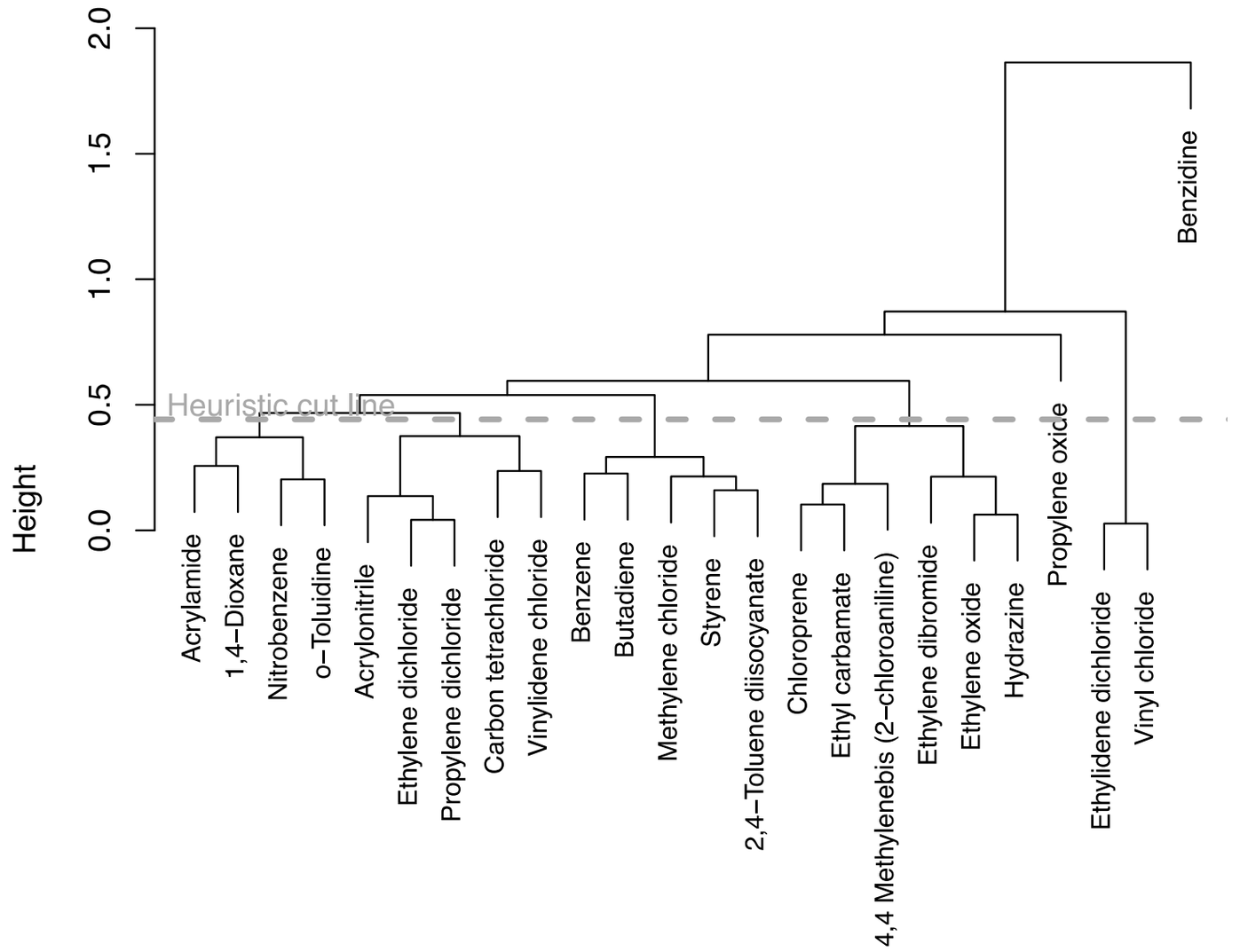
Our basic approach…

1. Use cluster analysis to find sets of MGCs with similar exposure patterns
   - Clusters may overlap
   - Patterns may use outcome
2. Choose “best” clustering
   - How? currently heuristic
3. Create summary measures based on “best” clustering
4. Evaluate risks associated with exposure summaries



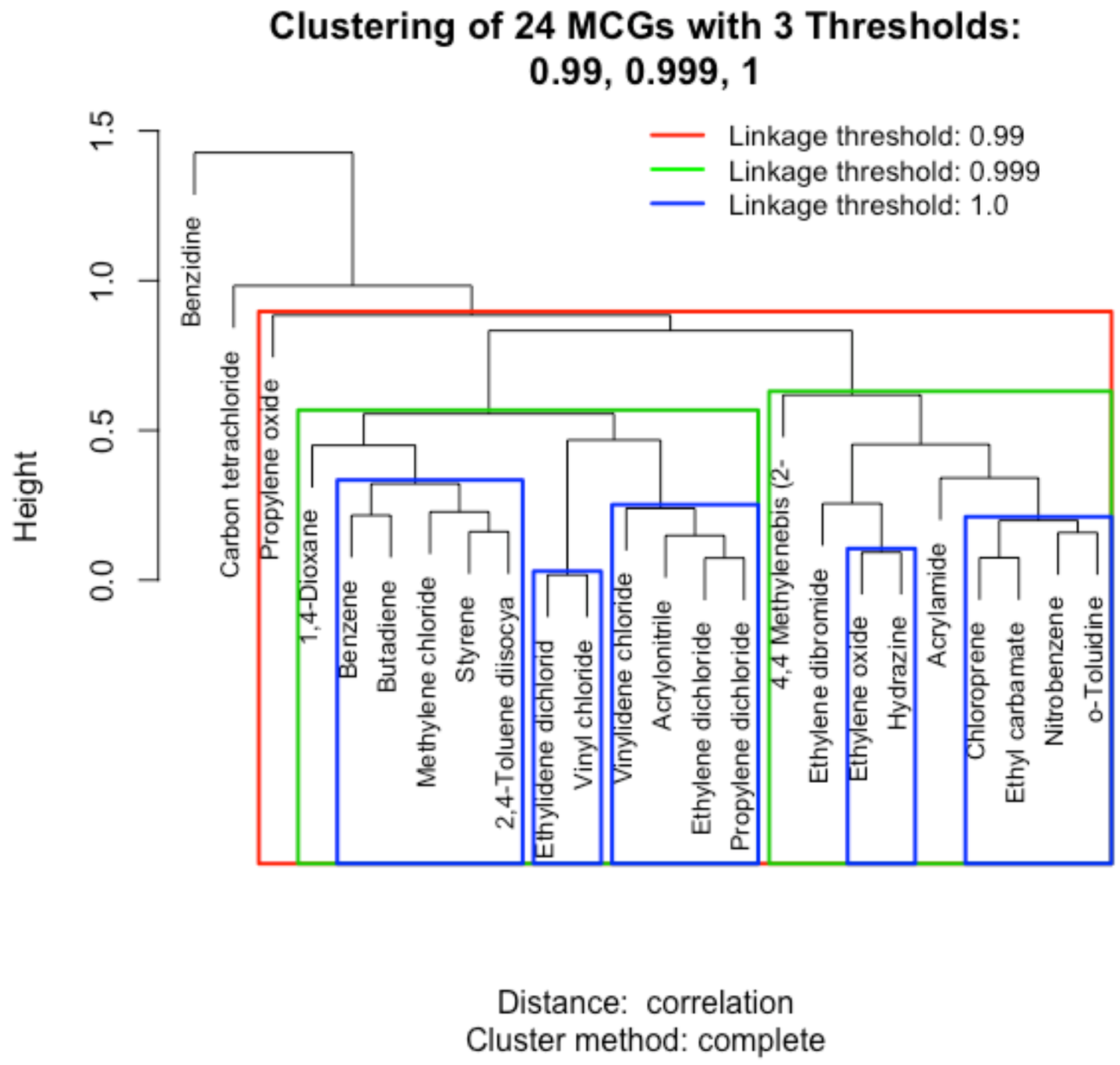
First way to cluster: traditional cluster analysis

- Use 1-Spearman correlation matrix as dissimilarity measure
  - Also exploring machine learning and graphical models
- Resulting **complete linkage** dendogram shows some clear clustering
- Big question: **where to cut the tree?**
  - Heuristic vs. computational



One approach to cluster choice: bootstrapping

- We use the R package ***pvclust***
  - Assesses uncertainty in hierarchical clustering
- Tests hypothesis that "cluster does not exist" with probability *p*
  - Collapse nested clusters that pass threshold
- "Linkage threshold" is just *(1-p)*
  - We look at a linkage threshold of 1



Second way to cluster: “supervised clustering”

- Borrowed from the “expression array” literature in genomics

- Their problem, in its simplest form:
  - Subjects classified into two classes, or “outcomes”
  - For each subject, have expression levels for ***lots*** of genes
  - Want to know what groups of genes predict outcome

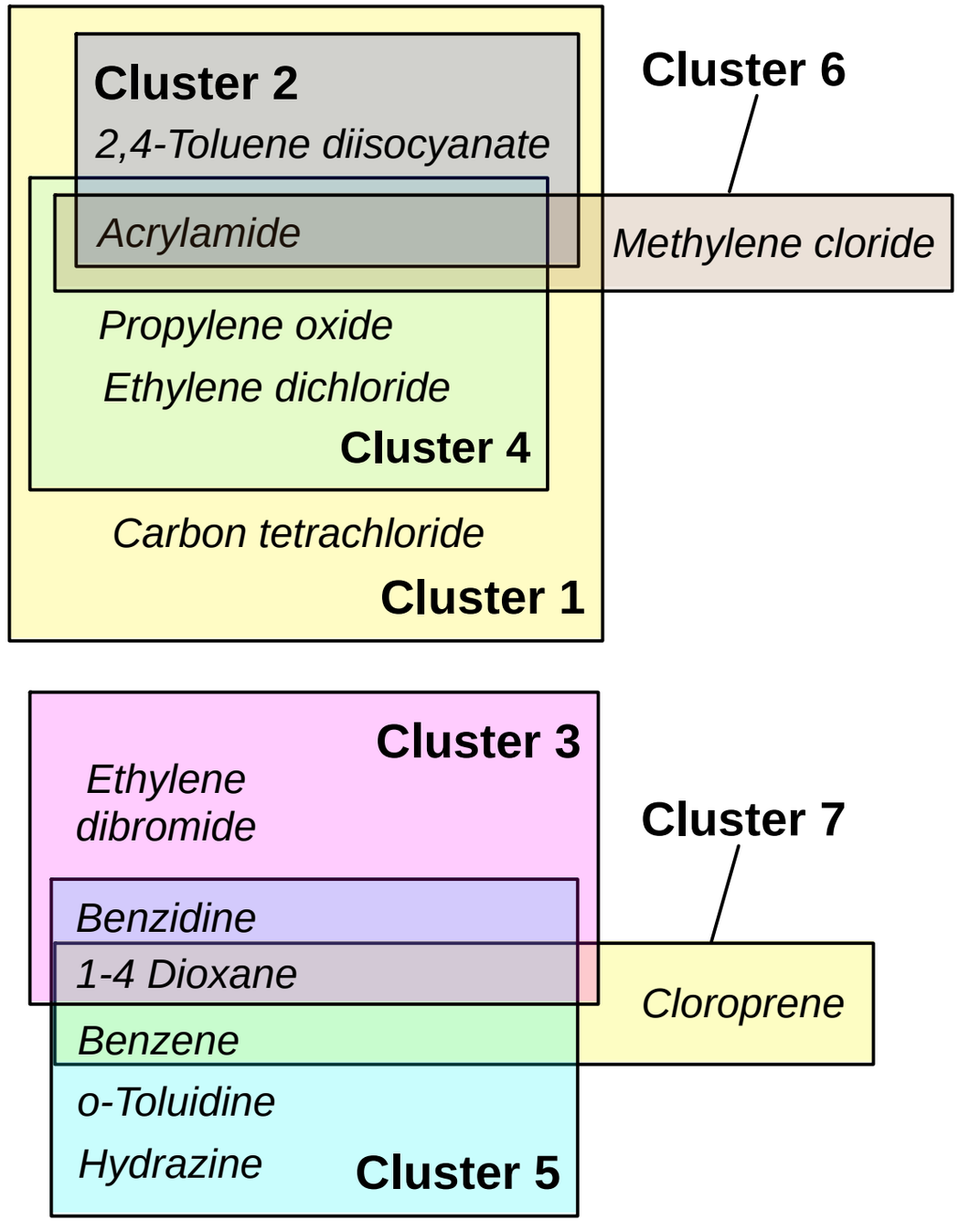
- Their complications (just like ours)
  - Genes hardly ever work alone: genes act along pathways
  - Only some pathways affect outcome
  - Some genes may be involved in multiple pathways

We used a program called “pelora”

- Stands for ***Pe***nalized ***Lo***gistic ***R***egression ***A***nalysis
- Uses exposure-outcome relationships to guide clustering
- Supervised approach to clustering that simultaneously combines
  - variable selection
  - variable grouping
  - sample classification
- In R package supclust, created by Dettling and Buhlmann
- Number of clusters is user defined
  - We looked at 2, 3, …, 10 clusters
- We chose 7 clusters
  - No new MGC entered a cluster for more than 7 clusters

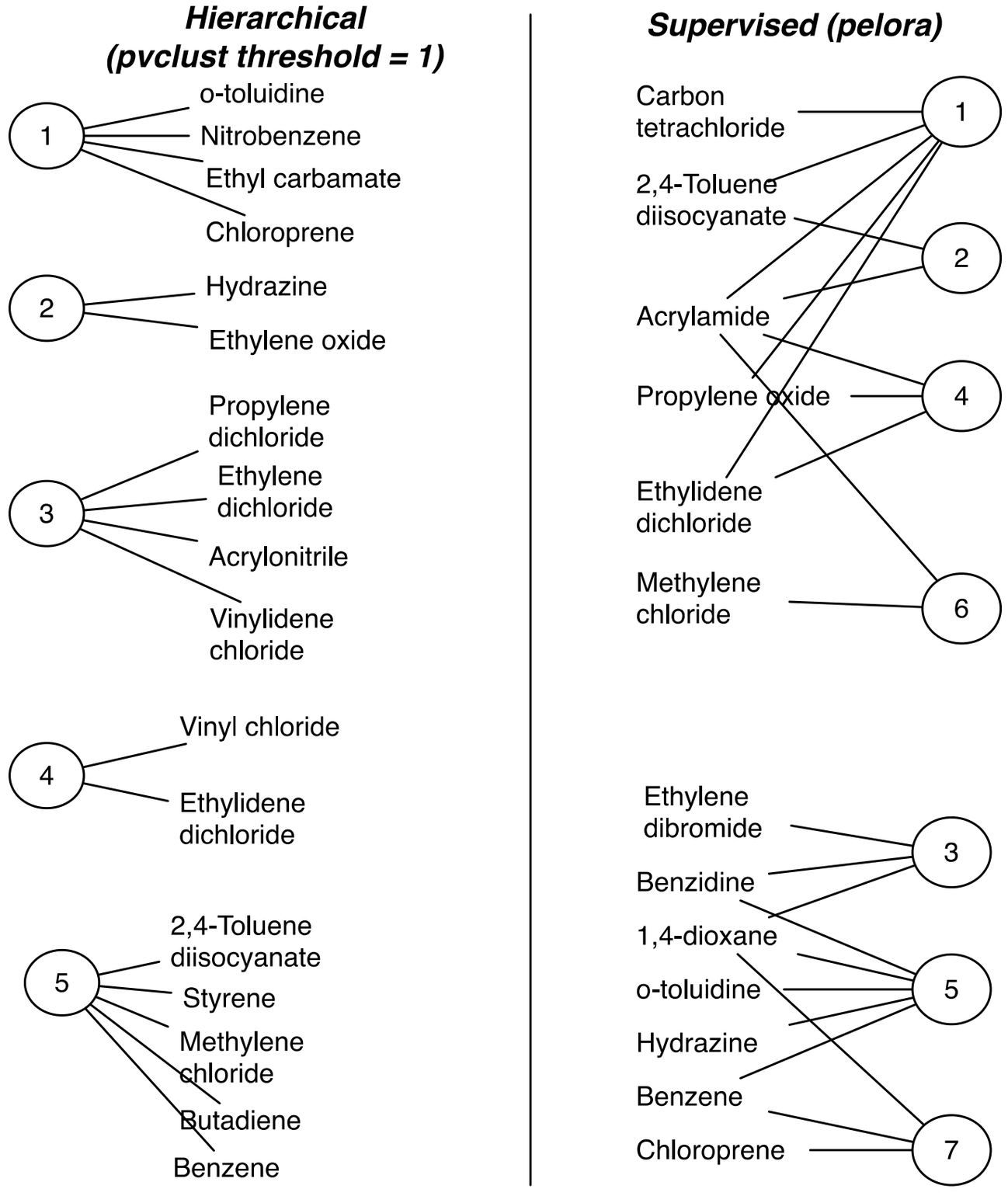
Our supervised clustering approach resulted with a set of seven ***overlapping*** clusters

- Two major groups
  - Clusters 1, 2, 4, 6
  - Clusters 3, 5, 7
- Not all compounds are in a cluster
  - Only 14 of 24 MGCs appear to affect outcome
- Acrylamide is in ***all*** clusters in group 1
- 1-4 Dioxane is in ***all*** clusters in group 2



Comparing cluster results from the two approaches: hierarchical vs. supervised

- Very little correspondence between
  - Clusters based on correlation and
  - Clusters based on effect



HAPs Clusters and Breast Cancer Risk

- No evidence of risk associations from the unsupervised clusters
- Supervised clustering detected a few “significant” clusters at $p<=0.05$
  - Significance depended on number of quantiles into which the concentration was cut
- Size and significance of risk depended on
  - How clusters were created
    - Supervised vs. unsupervised
  - How “best” clustering was chosen
    - Bootstrapping for unsupervised clustering
    - Number of clusters in “best” clustering for supervised
      - Here we were quite heuristic
  - Number of clusters in final clustering

Conclusions

- Two methods of clustering produced quite different, but complementary results
  - They each produced interesting and potentially important summary measures
  - But clustering results were sensitive to the way the data were modeled
    - Another example of “Rashomon effect”

- Important to remember that high correlation between MGC concentrations in the environment may not correspond to similar underlying mechanistic properties

Summary

- Real world environmental exposures do not occur one chemical at a time
- There are many ways to try to account for mixtures of chemical exposures
- Our empirical analyses suggest the importance of assessing clustering in the context of outcome
  - BUT important to keep in mind that there is not one single answer
- This example was limited to HAPs, but there are many more chemical exposure opportunities out there
- We still have a lot to learn -- this is a topic in need of transdisciplinary work