Exploratory Analysis of Clinical Safety Data to Detect Safety Signals

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville TN USA
f.harrell@vanderbilt.edu
biostat.mc.vanderbilt.edu

Slides and R Code at Jump: FHHandouts

FDA Visiting Professor Lecture Series

White Oak MD 11 Oct 2005

Outline

1. The problem
2. Statistical efficiency issues
3. Graphs, not tables
4. ECDFs and extended box plots
5. Clinical trial data
6. Empirical CDFs for lab variables
7. Variable clustering
8. Time trends and clustering of AEs
9. Clustering of lab variables
10. Recursive partitioning
11. Who is having selected AEs?
12. Which AEs and lab abnormalities are independently related to treatment?
13. Examples from other trials

The Problem

- No analytic plan
- Often collect safety data to learn what safety parameters are of concern
- Large number of lab parameters and types of adverse events encountered
- More emphasis on type II error vs. type I error in contrast to efficacy assessment
- Adjustment for multiple P-values and ad hoc comparisons ill-defined; done informally but with an eye on consistent patterns
- Exploratory multivariate problem

Statistical Efficiency Issues

- For clinical lab data, common to compute proportions of subjects above $2\times$ or $3\times$ upper limit of normal
- “Normal” is arbitrary, not tied to clinical outcomes
- Most efficient cutoff of a continuous variable is the median
- Still results in efficiency of only $\frac{2}{\pi} \approx 0.64$ if distribution is normal (median test)
- Wilcoxon 2-sample test has efficiency of $\frac{3}{\pi} \approx 0.95$
- Nonparametric tests generally have greater efficiency than parametric tests because of skewness, high-influence points, etc.

Efficiency Issues, *continued*

- Keep continuous variables continuous as much as possible
  - Simple tests: Wilcoxon, Kruskal-Wallis, Spearman
  - Graphs: empirical cumulative distributions, scatterplots, multiple quantiles

Graphs, Not Tables

- Have pity on statistical and medical reviewers
- Difficult to see patterns in tables
- Substituting graphs for tables increases efficiency of review

ECDFs and Extended Box Plots

- Empirical cumulative distribution functions:
  - full resolution of data
  - unique; no arbitrary binning of data
  - ideal for comparing 2 treatments
- Box plots can be extended to show not only 3 quartiles but other quantiles [2]
- 0.25, 0.5, 0.75, 0.9 intervals + median and mean

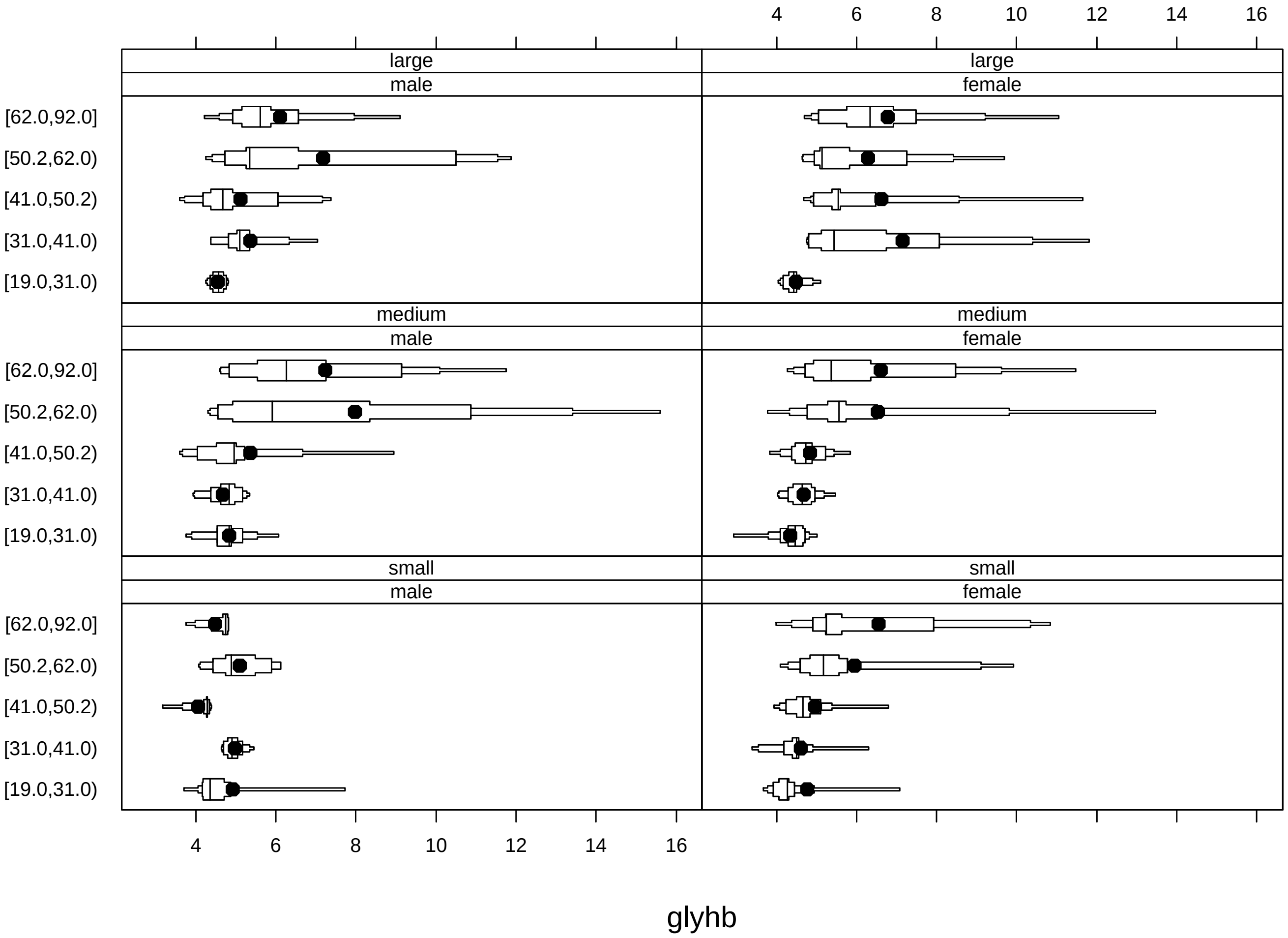

Figure 1: Extended box plots for glycosolated hemoglobin, stratified by two categorical variables (forming panels) and one continuous variable (categorized into quintiles).

Clinical Trial Data

- A pharmaceutical company generously supplied excellent demographic, AE, vital signs, clinical chemistry, and ECG data
- Three protocols combined
- Phase III randomized double-masked placebo-controlled parallel-group studies
- Drug:placebo 2:1 randomization ($n = 1374$ and 684)
- Analyzed asessments at weeks 0, 2, 4, 8, 12, 16, 20 (plus week 1 for AEs)

Comparing Lab Variables Between Groups

- Means and SDs are not very helpful for highly skewed data
- Examining summary stats individually can exaggerate treatment differences
- Empirical CDFs display all information objectively

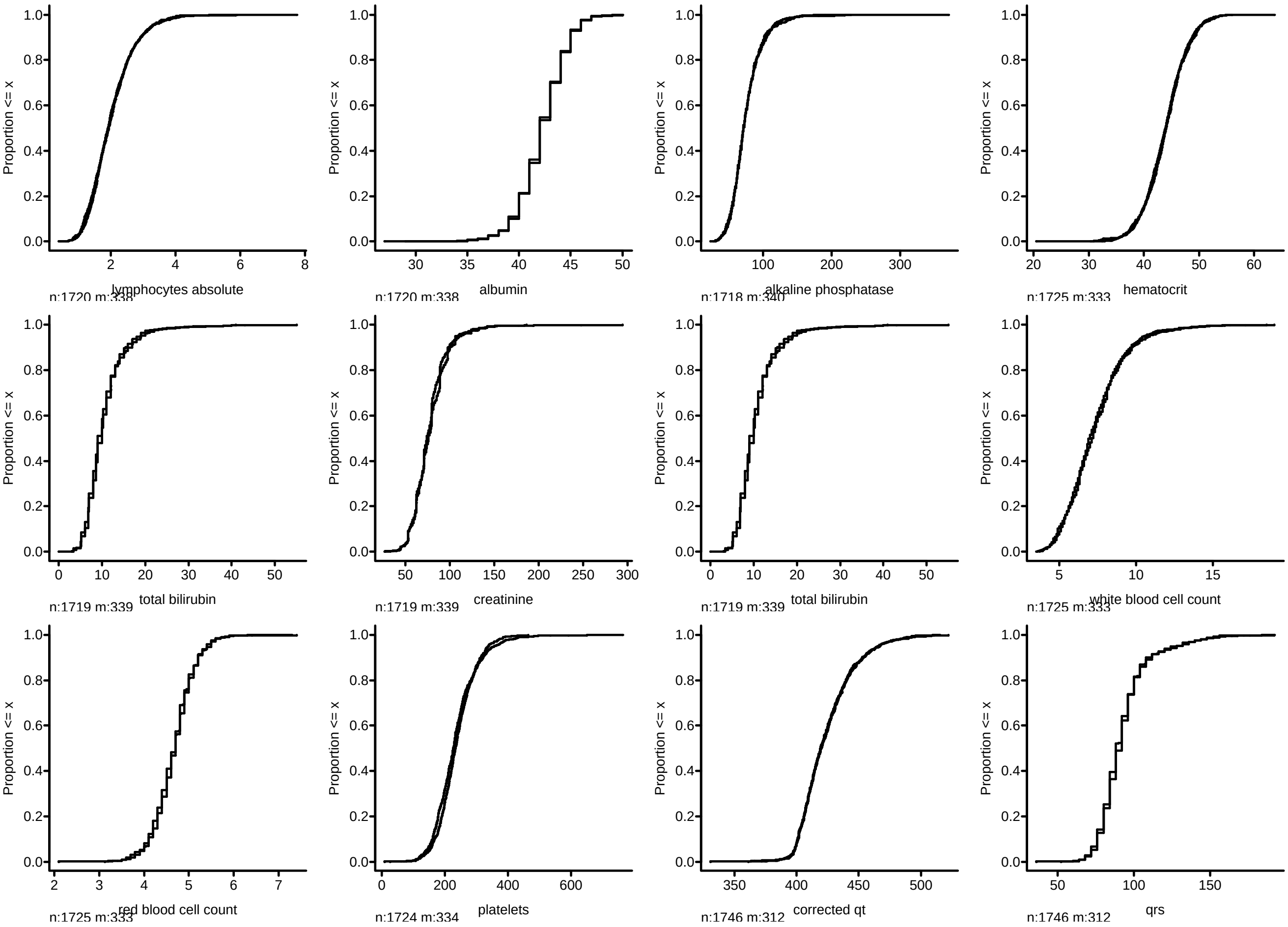

Figure 2: Empirical CDFs of 12 lab and ECG parameters stratified by treatment group for week 8. CDFs are virtually superimposed.

Usefulness of Variable Clustering

- Learn how multivariate responses occur/move together
- Learn about redundancy—variables that are sufficiently captured by other variables
- Cluster separately by treatment to see if treatment more likely than placebo to cause multiple abnormalities in the same subject
- Standard hierarchical clustering algorithms may be run on a variety of similarity matrices based on pairwise similarity measures
  - proportion of subjects missing on two variables
  - proportion of subjects having a pair or AEs
  - Spearman ρ^2: strength of monotonic relationship

Clustering of AEs

- Analyzed 7 AEs having at least 100 episodes
- Which AEs occur together?
- Similarity measure: proportion of patients having both AEs (diagonal = 1)

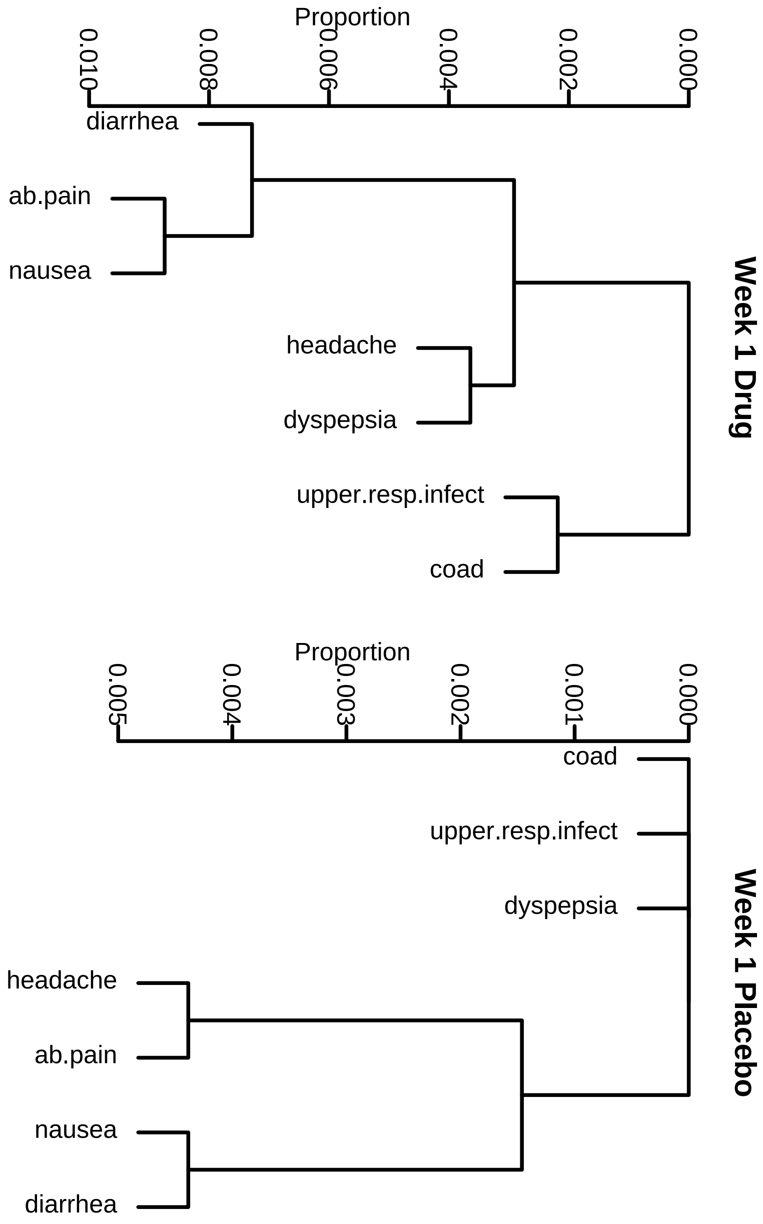

Figure 3: Variable clustering of AEs at week 1 using proportion of patients having two AEs as similarity measure.

Time Trends of AE Clustering

- Difficult to track changes in dendograms over multiple times
- Can separately plot all pairwise similarities over time, stratified by treatment
- Estimate incidence of pairs of AEs above coincidence levels
- $P_{ij} - P_i P_j$

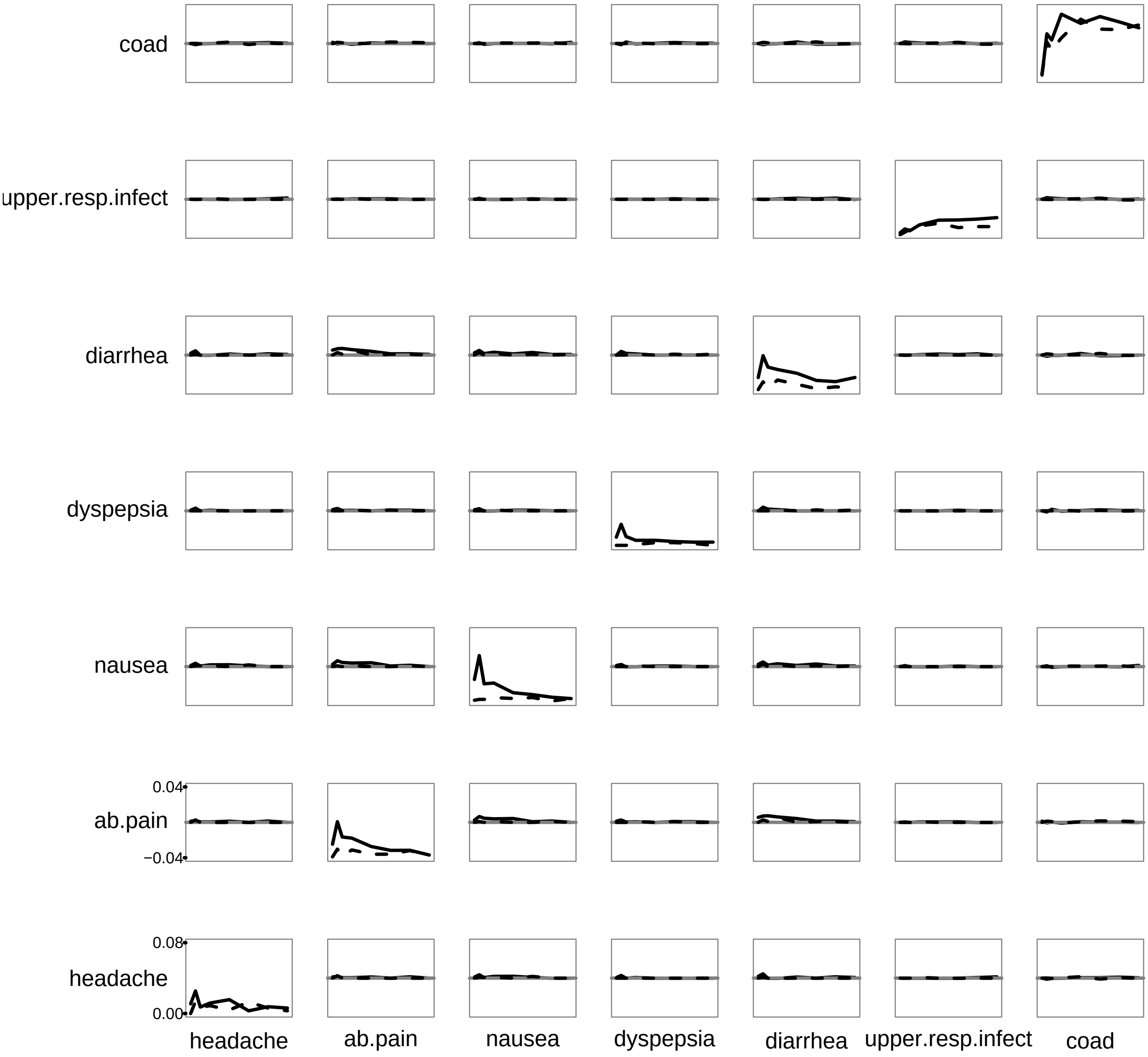

Figure 4: Time trends in incidence of AEs (diagonal) and chance-corrected joint incidence (off-diagonal). Solid lines represent drug and dotted lines placebo. Horizontal reference lines are at zero (chance level of joint incidence). Week is on x-axes.

Clustering Lab Variables

Similarity measure = Spearman ρ^2

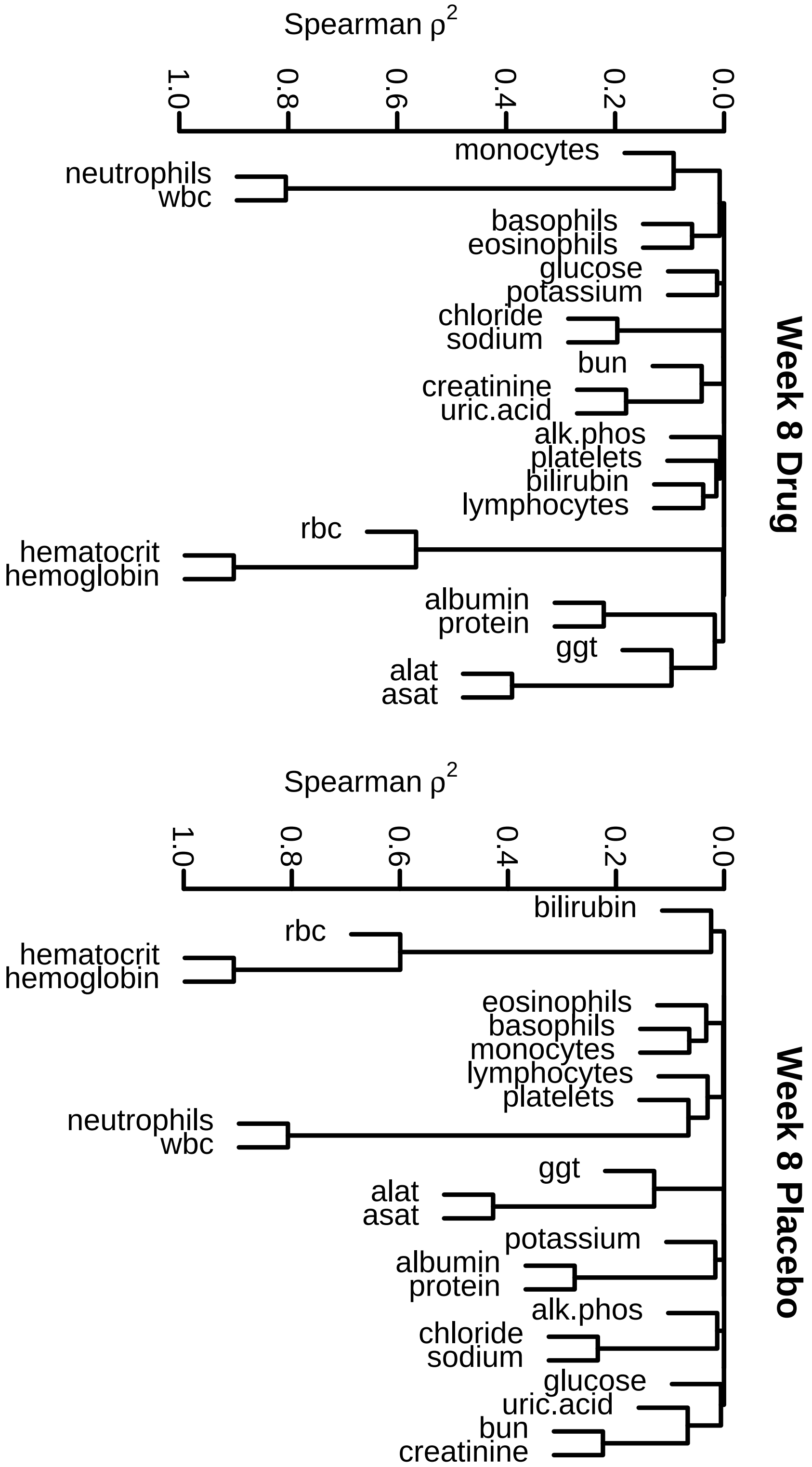

Figure 5: Variable clustering of clinical chemistry variables at week 8.

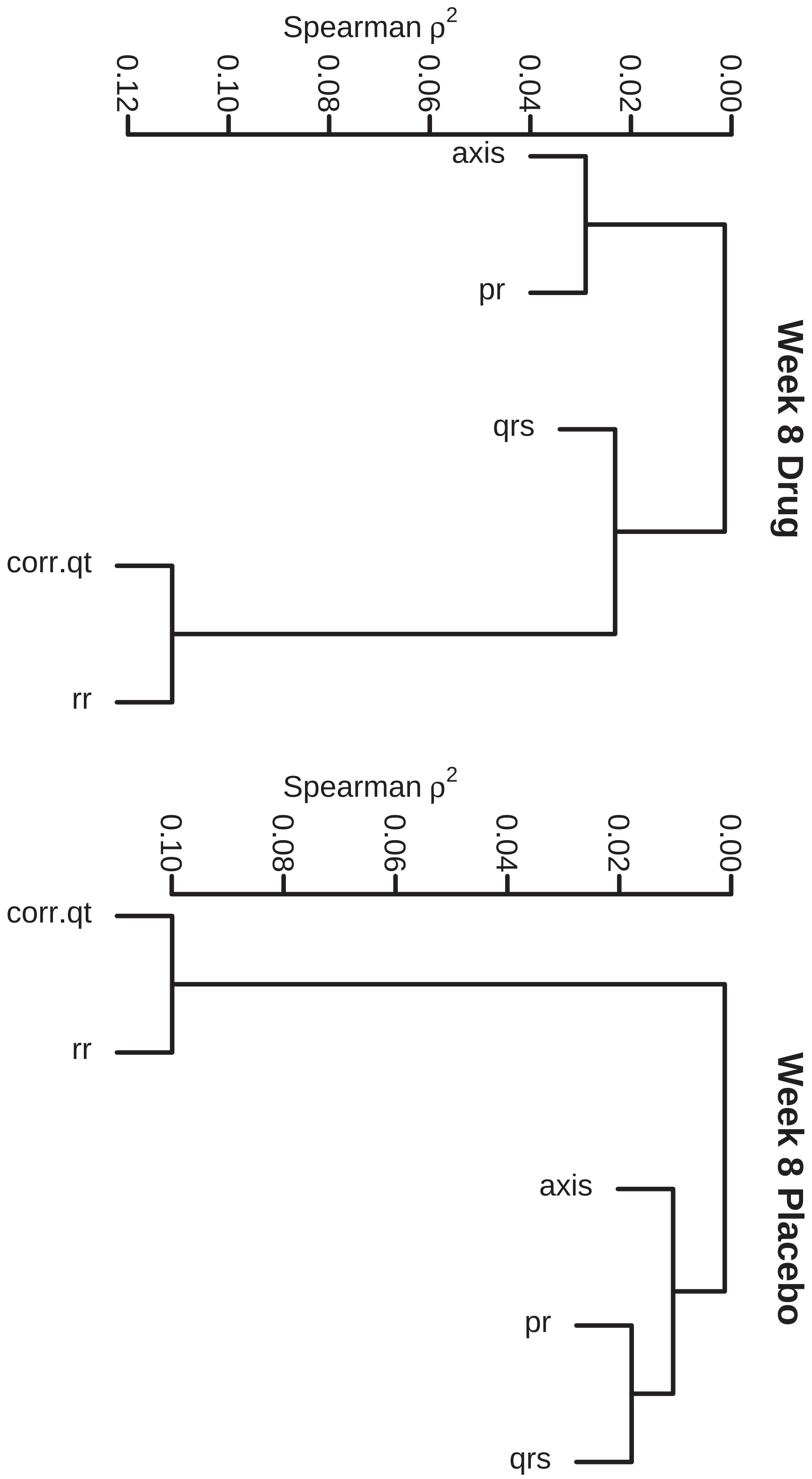

Figure 6: Variable clustering of ECG parameters at week 8.

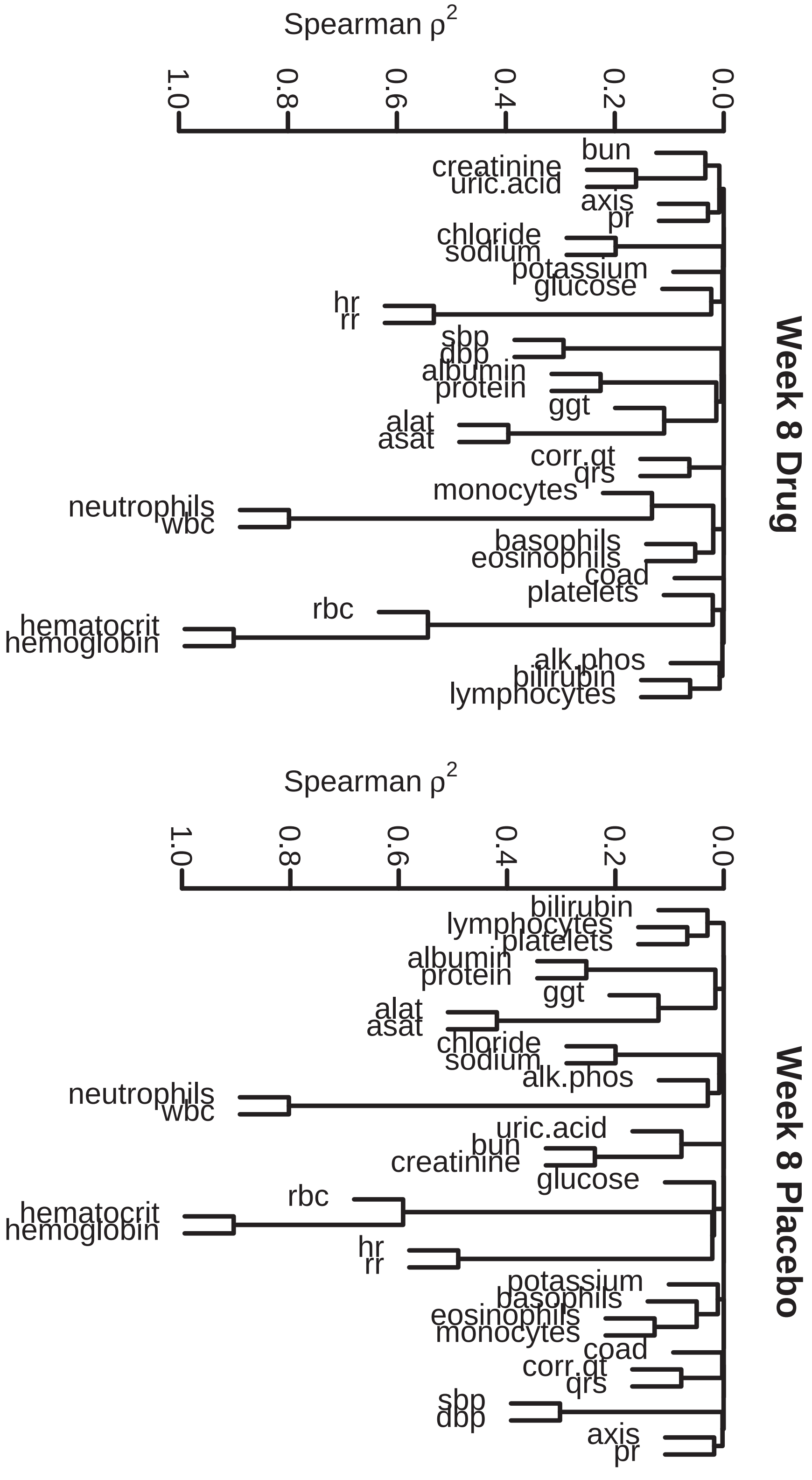

Figure 7: Variable clustering of combined vital signs, AE, and clinical chemistry variables at week 8.

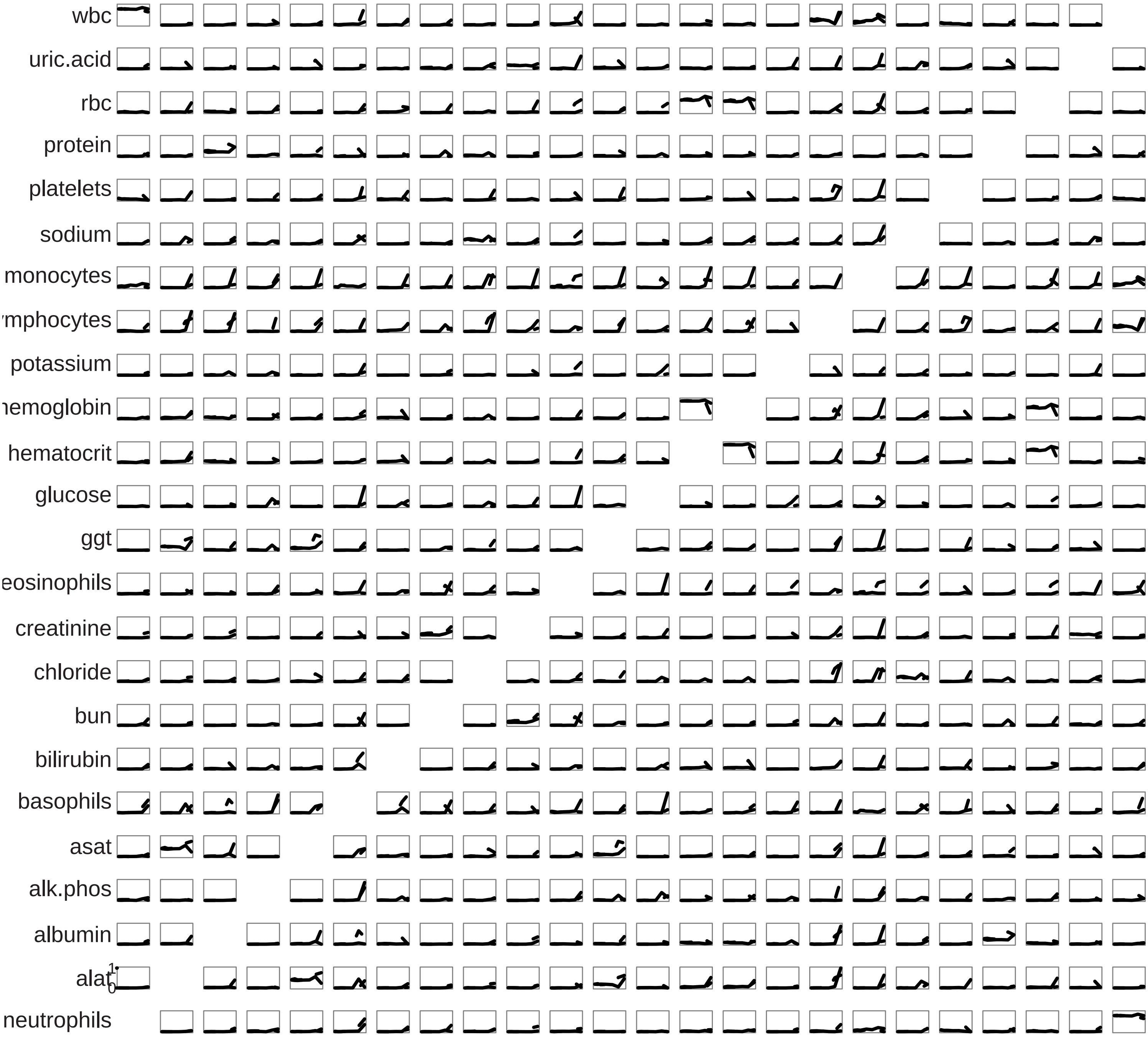

Figure 8: Time trends in correlation between selected lab variables, stratified by treatment (dotted line = placebo). Y-axes are Spearman ρ^2. Week is on all x-axes.

Summarizing Time Trends in Correlations

- For each treatment compute slope of squared rank correlations over time
- Compute differences (drug - placebo)

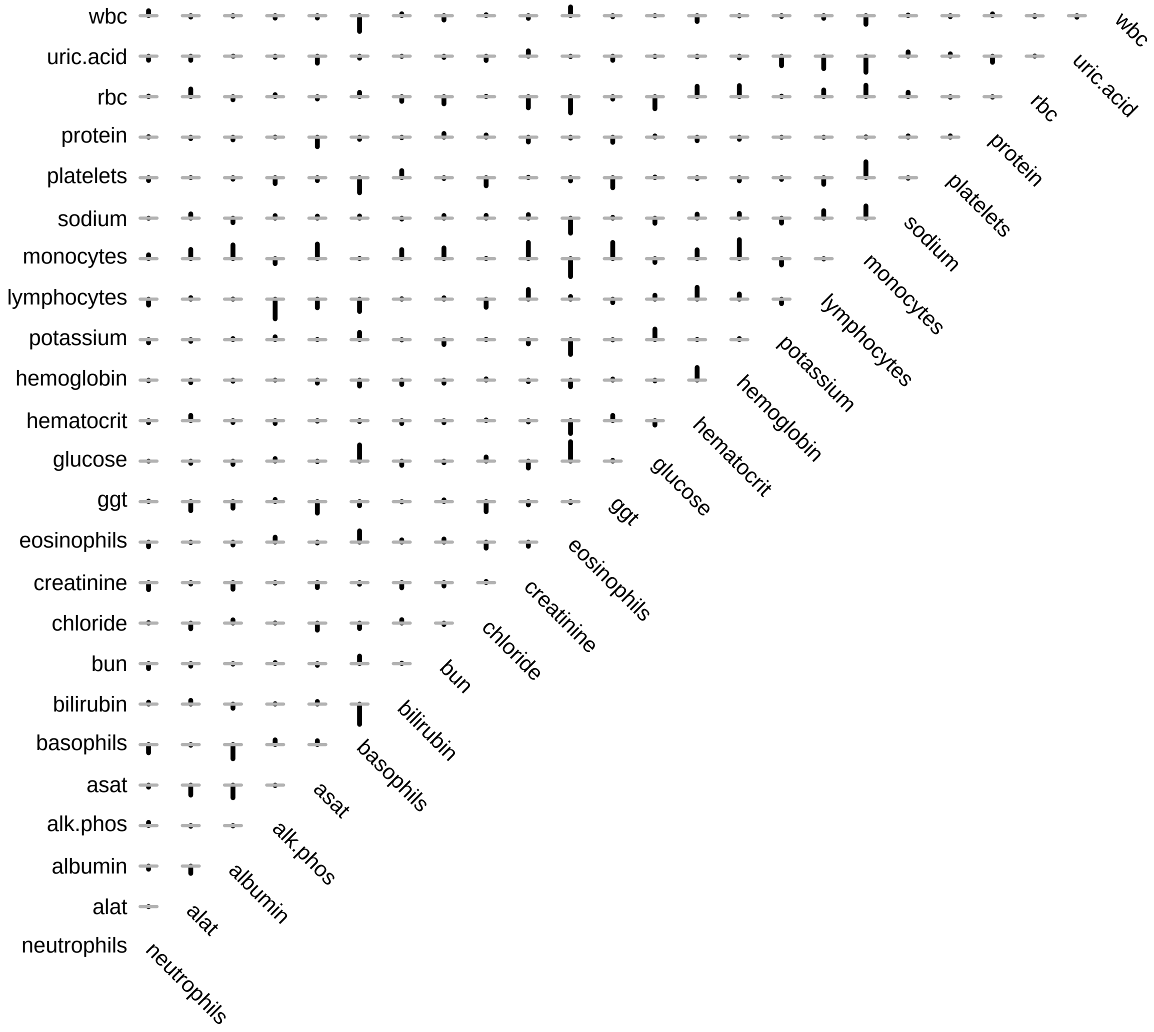

Figure 9: Differences in slopes of squared rank correlations over time. Vertical line segments above gray horizontal reference lines correspond to positive slope differences and hence indicate that correlations became stronger over time for drug than for placebo; those below the gray lines correspond to negative slope differences and hence indicate that the rank correlation between the two indicated variables became smaller over time for drug as compared to placebo. Maximum difference was 0.033 and minimum was -0.035.

Recursive Partitioning

- Almost model-free
- Useful for handling large numbers of potential predictors
- Can provide interesting leads if not internally validated
- Conservative if tree pruned to internally validate

Who is Having an AE?

- Chronic obstructive airways disease is the most common AE
- Use recursive partitioning to develop a regression tree predicting Prob(COAD)
- Descriptive analysis; requires validation
- Candidate predictors: treatment, time, 6 demographics, 2 smoking, systolic and diastolic bp, 24 clinical chemistry parameters, 5 ECG parameters

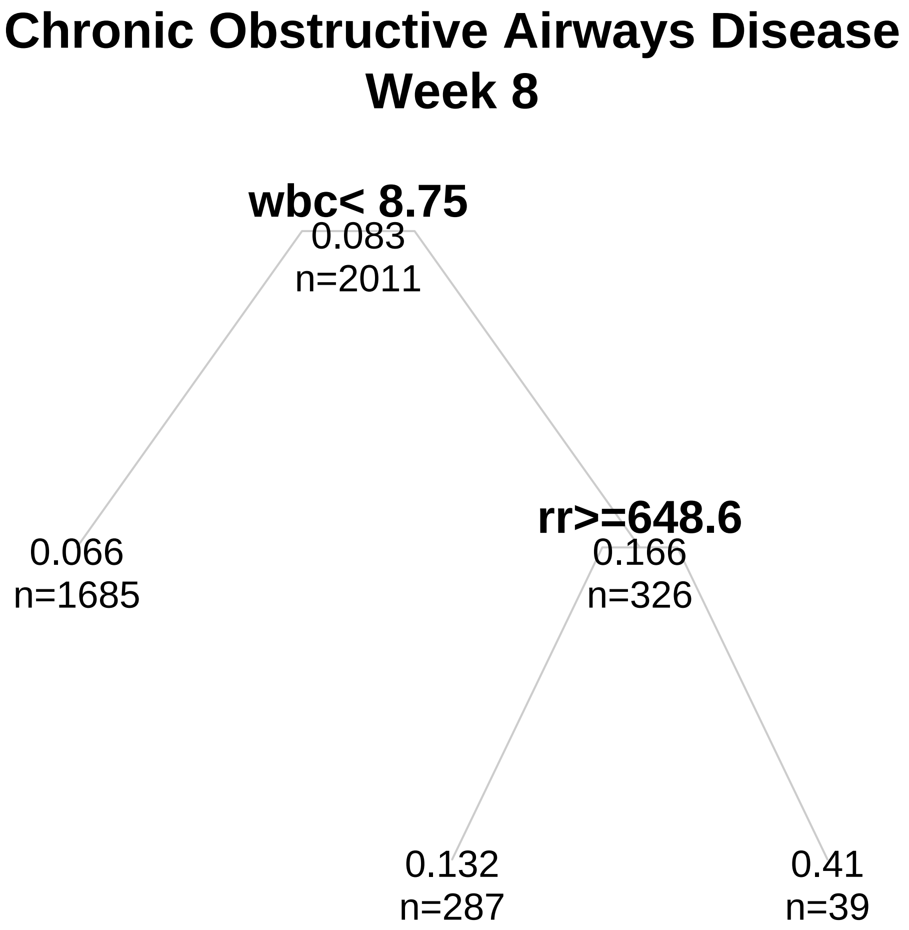

Figure 10: Regression tree predicting Prob(COAD) at week 8.

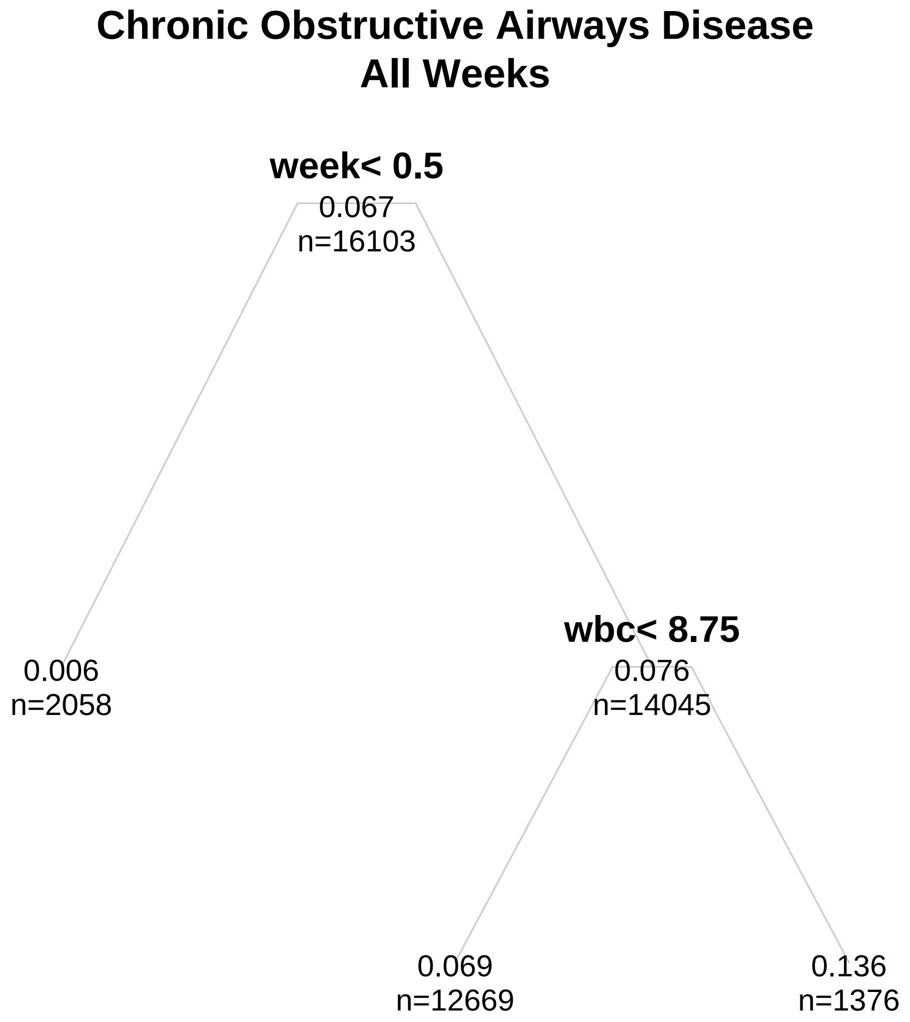

Figure 11: Regression tree predicting Prob(COAD) at any week.

Logistic Model Predicting COAD

- Predict COAD at week 8 for subjects having no COAD before week 8
- Predictors: treatment, age, sex, smoking, restricted cubic splines in lymphocytes, WBC, heart rate, BMI
- Only 54 events and 458 non-events due to missing data (especially lymphocytes)
- Recursive partitioning using treatment, baseline variables, vital signs, EKG variables, and clinical chemistry variables could not find a split that validated
- Logistic model: $P = 0.11$ for treatment, 0.04 for lymphocytes (very nonlinear), 0.06 for WBC (nonlinear)
- $C = 0.739$ (apparent) 0.66 (bootstrap overfitting-corrected)

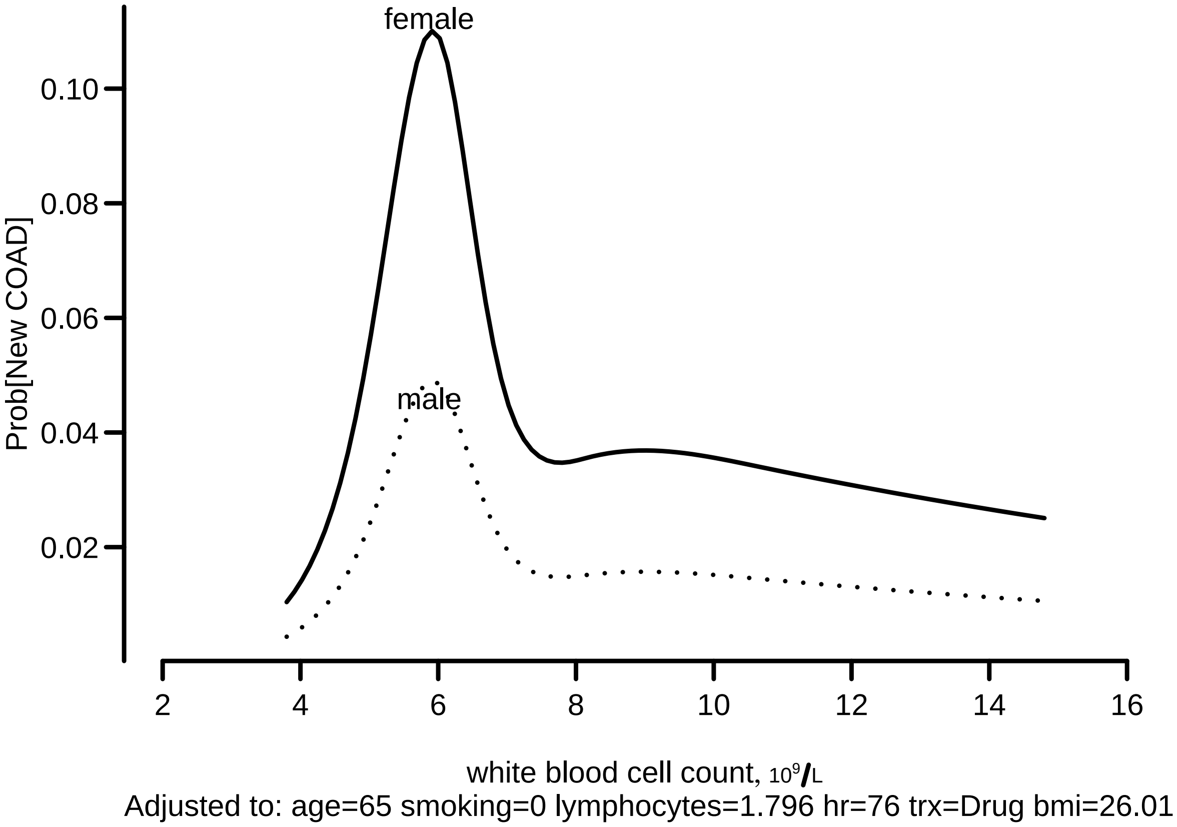

Figure 12: Predicted probability of new COAD as a function of WBC and sex from binary logistic model.

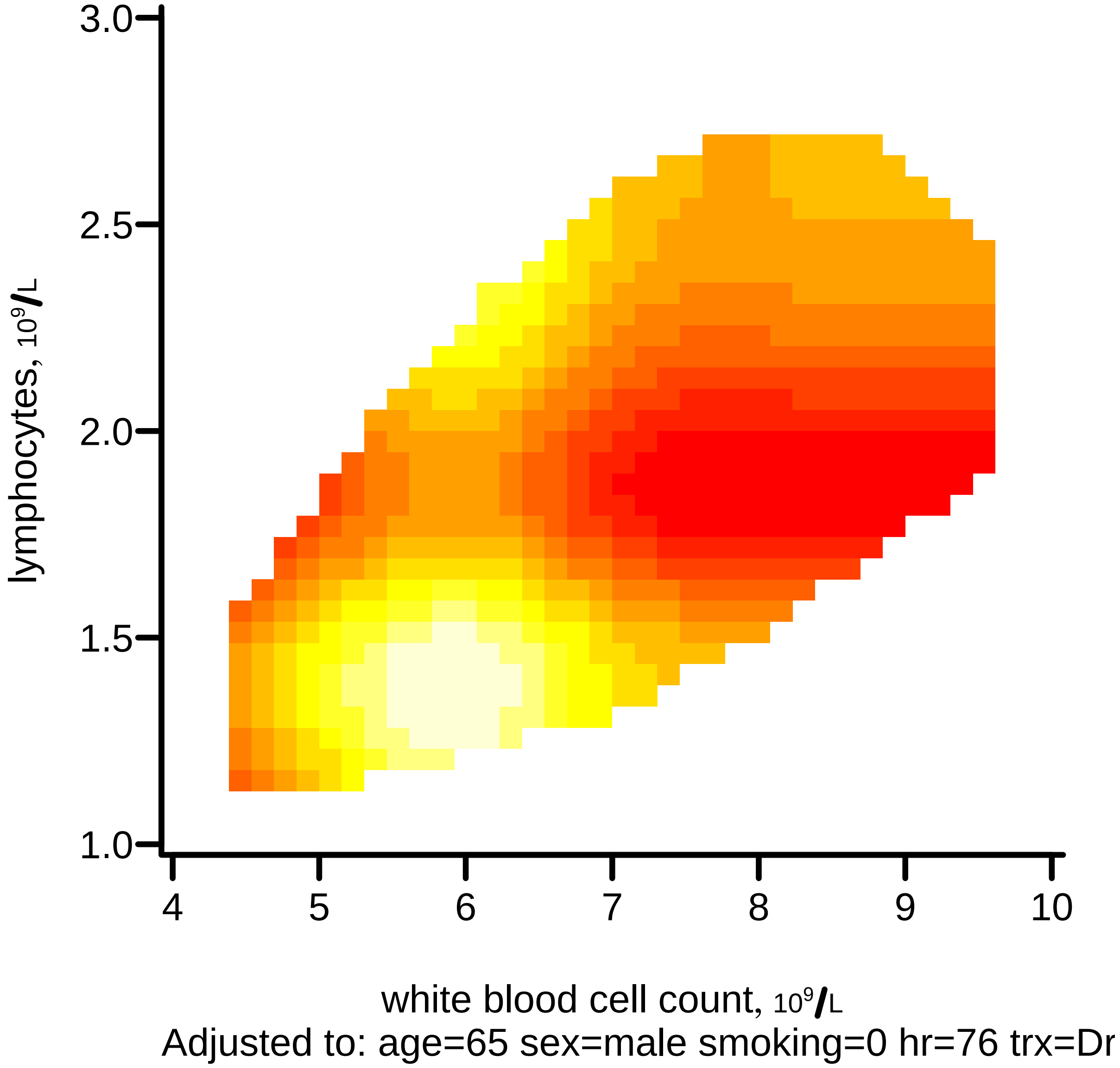

Figure 13: Predicted probability of new COAD as a function of WBC and lymphocytes. Regions beyond which fewer than 10 subjects exist are not shown.

Prediction of GI Problems

- Recursive partitioning to predict union of abdominal pain, nausea, dyspepsia, diarrhea at 8 weeks using predictors measured at 4 weeks (except AEs) plus some baseline variables.
- Used treatment, baseline variables, vitals, EKG, clinical chemistry
- No splits that cross-validated
- Try predictors treatment, age, sex, BMI, smoking, SBP, DBP, HR, WBC
- Again no splits
- Logistic regression using same variables (with splines) — using 161 GI events, 1577 non-events
- Apparent C=0.64, bootstrap validated 0.58
- Males less likely to have GI AE ($OR = 0.58, P = 0.004$)

- Treatment: $P = 0.15$
- Low SBP associated with $\uparrow$ GI events

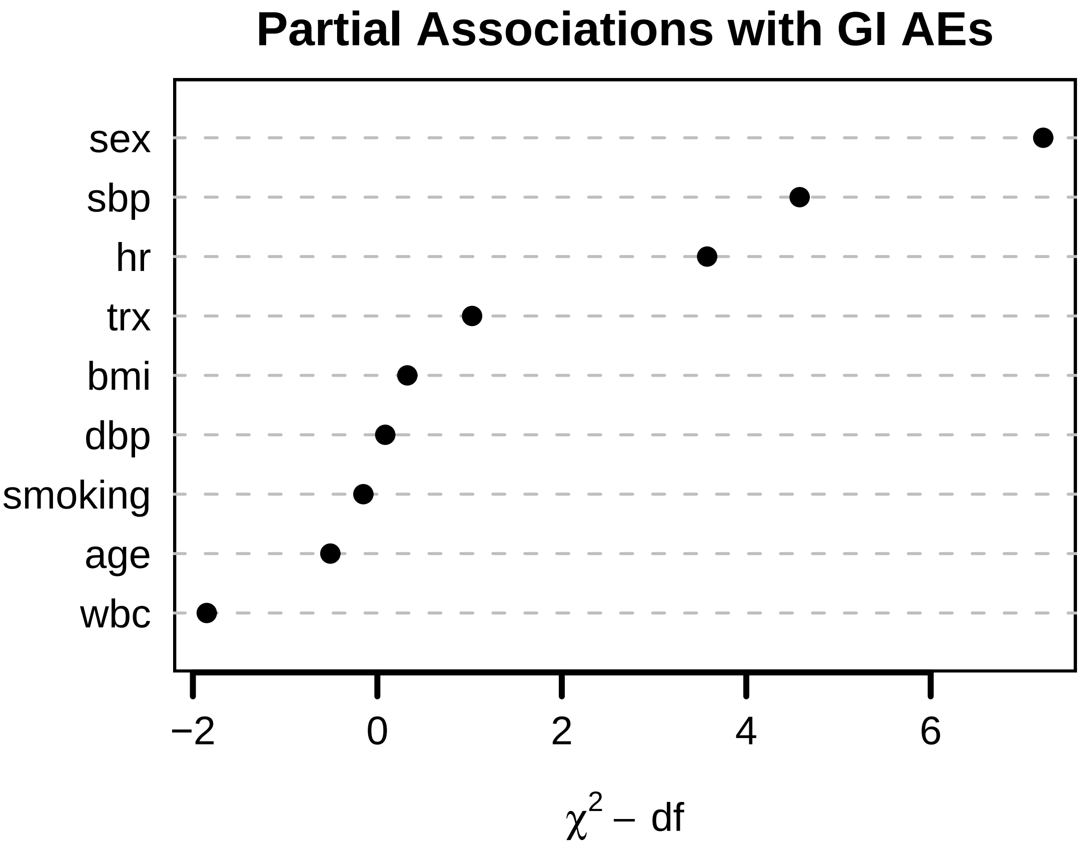

Figure 14: Tests of partial association of subject characteristics at week 4 or baseline, with GI events at week 8. χ^2 values are adjusted for d.f.

Multivariate Analysis of Treatment Differences

- Multiple responses: AEs, vital signs, labs
- True multivariate methods are cumbersome and make many assumptions
- O'Brien [3] turned the 2-sample t-test backwards
- Predict treatment from Y using binary logistic model (propensity score [1])
- To allow differences in means and variances use Y, Y^2
- Extend to multiple Ys: more flexible than Hotelling T^2
- Start with recursive partitioning

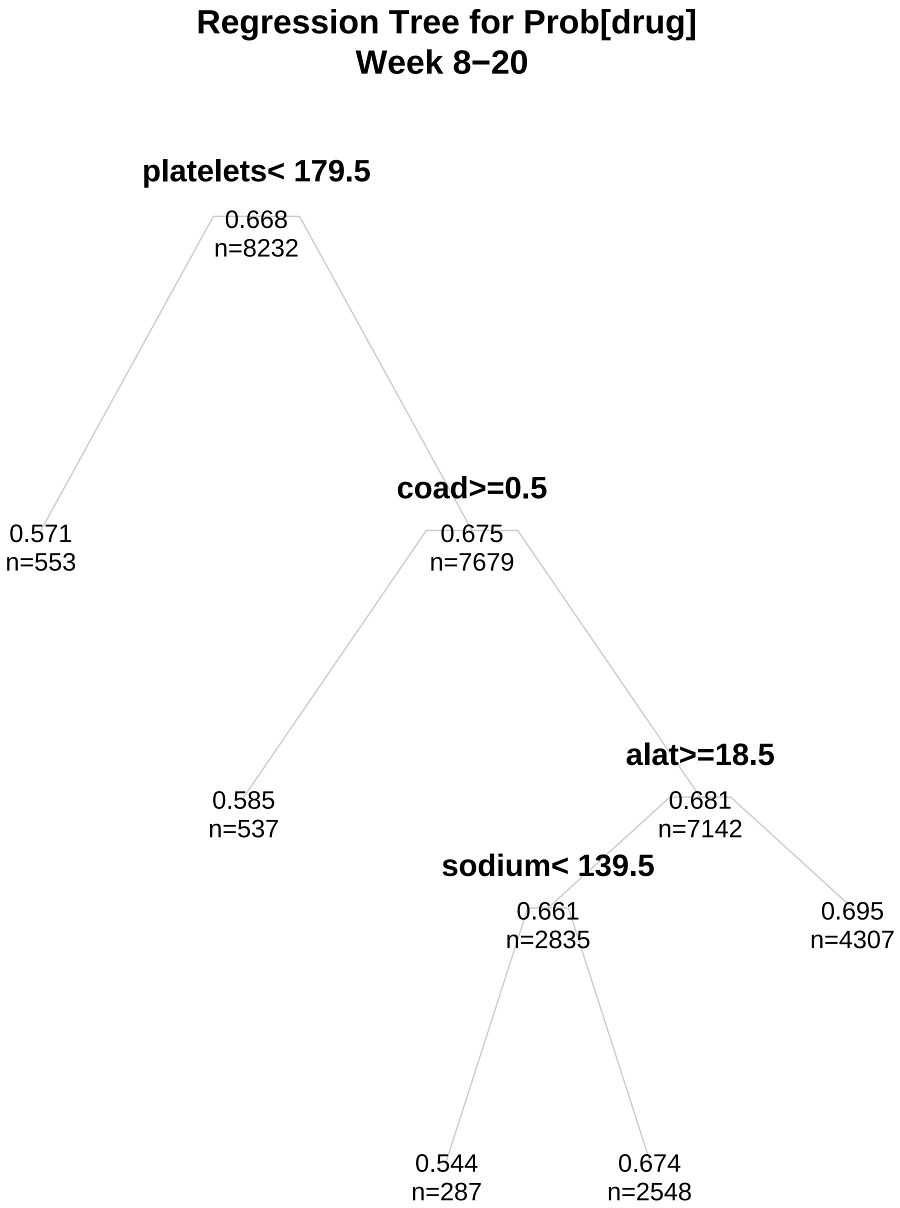

Figure 15: Regression tree predicting Prob(drug) using safety responses for weeks 8-20. When an inequality holds for a subject, branch to the left, otherwise to the right. `coad>=0.5` means that chronic obstructive airways disease is present. `alat` stands for alanine aminotransferase.

Estimating Prob(conditions $C|$treatment)

- Bayes' rule:
  $P(C|\text{drug}) = P(\text{drug}|C)P(C)/P(\text{drug})$ = $P(\text{drug}|C)P(C)/\frac{2}{3}$
- $P(C|\text{placebo}) = [1 - P(\text{drug}|C)]P(C)/\frac{1}{3}$
- $RR = P(C|\text{drug})/P(C|\text{placebo}) = \frac{1}{2}P(\text{drug}|C)/[1 - P(\text{drug}|C)]$
- Example: If $P(\text{drug}|C) = \frac{2}{3}$, drug:placebo RR of C = 1
- drug:placebo RR that platelets are below 180 = $\frac{1}{2}0.571/.429 = 0.67$
- drug:placebo RR that platelets are above 179 and COAD is present = $\frac{1}{2}.585/.415 = 0.71$.

Binary Logistic Model for Prob(drug)

- Assume additivity
- Do not assume linearity
- Restricted cubic splines for continuous variables
- Wald χ^2 for each variable gauges the partial association between that variable and treatment after adjusting for associations between all other variables and treatment

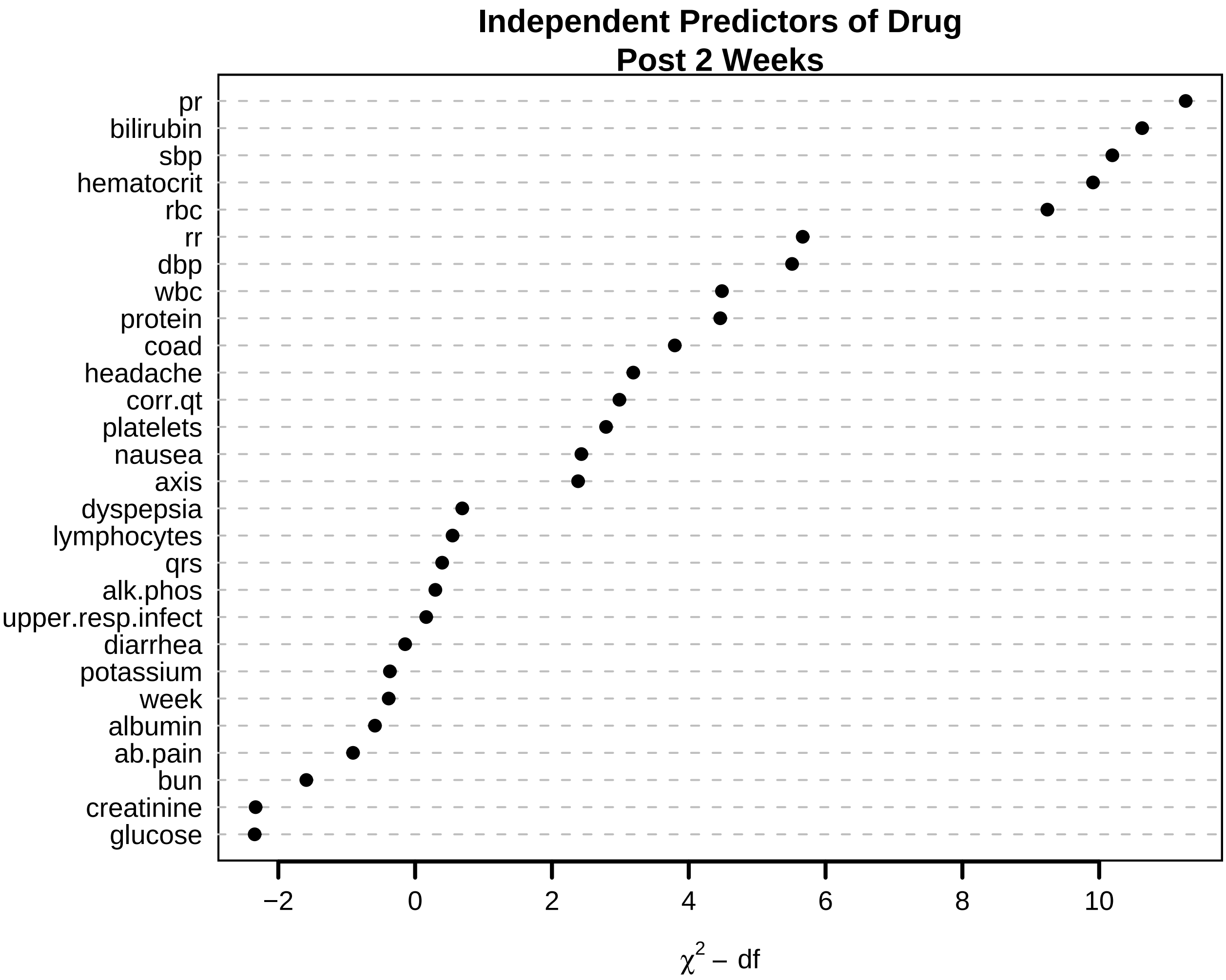

Figure 16: Degree of partial associations with treatment.

- Similar studies A (23 subjects) and C (49 subjects)
- Multiple doses and days (one dose/subject)

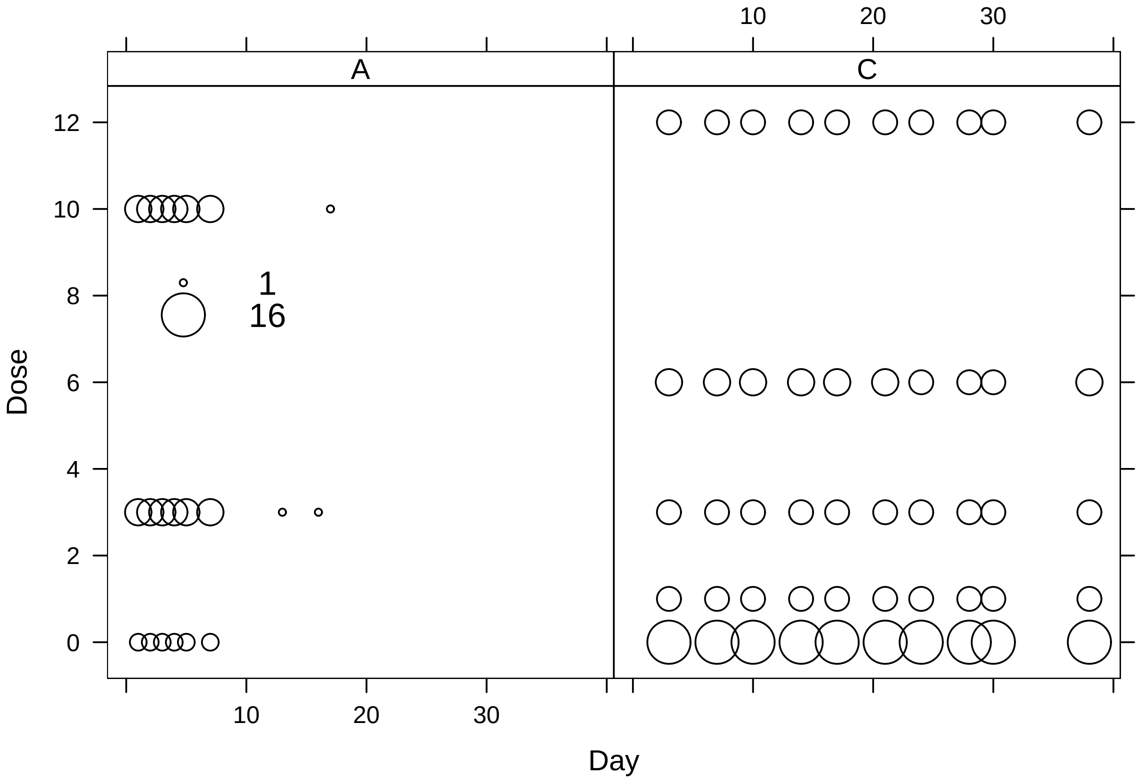

Figure 17: Study designs. Size of bubble is proportional to number of subjects (see key for 1 and 16 subjects). Left panel is study A, right is C.

Data

- 5 normalized liver function parameters

TB	Total Bilirubin
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
GGT	γ Glutamyl Transferase
AP	Alkaline Phosphatase

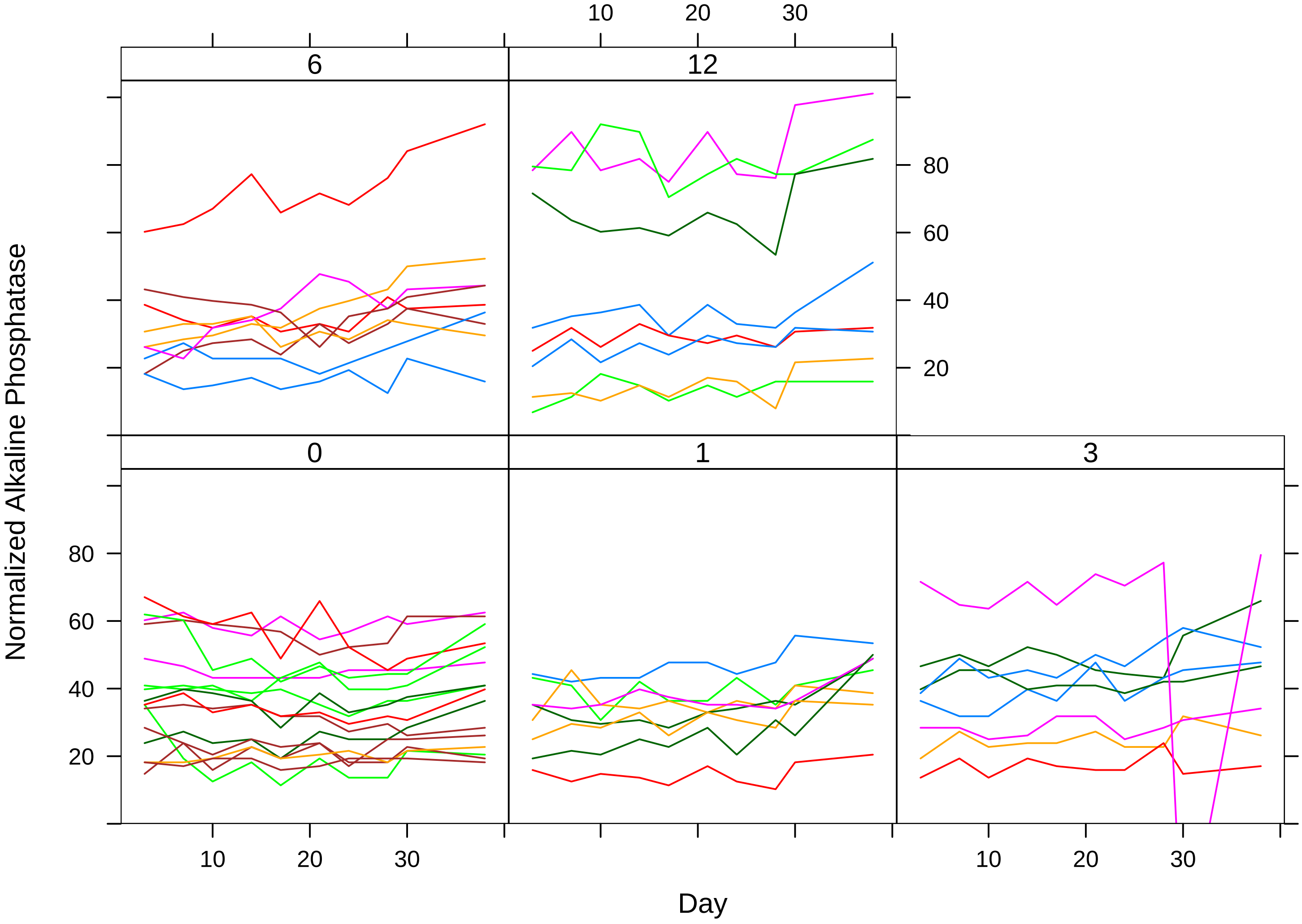

Figure 18: Individual normalized alkaline phosphatase measurements for study C, for 5 doses. Each line represents one subject.

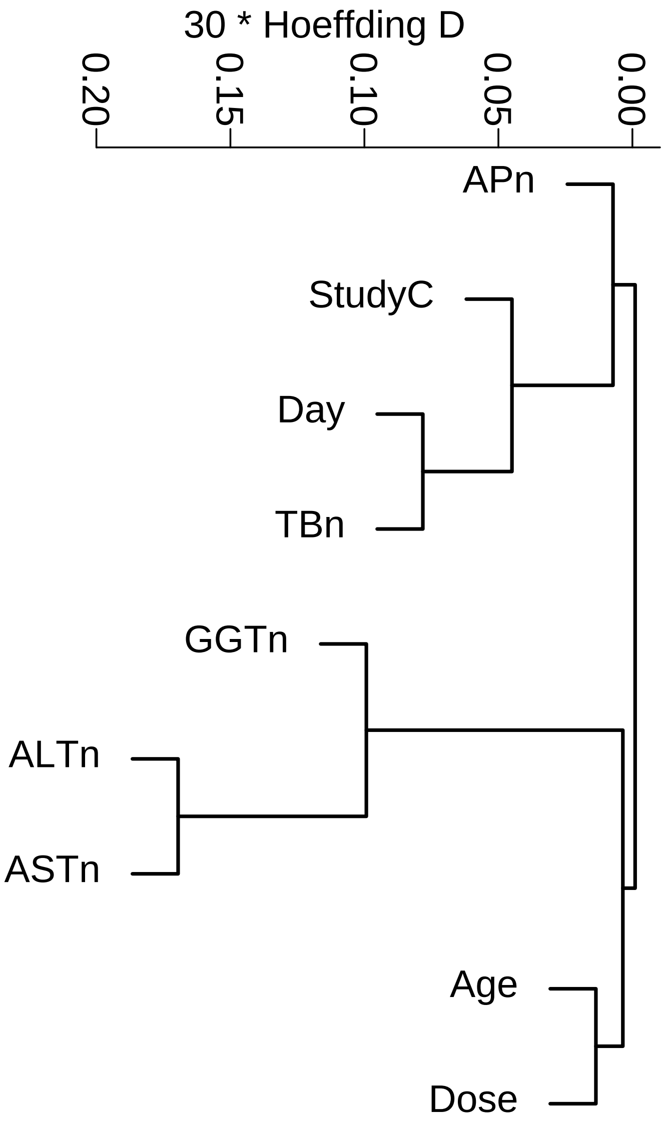

Figure 19: Clustering of individual normalized post-baseline clinical chemistry parameters using Hoeffding's D index as a similarity measure.

Data Reduction

- Data reduced to summary scores: within-subject slope of response vs. time, and AUC
- Slopes emphasized

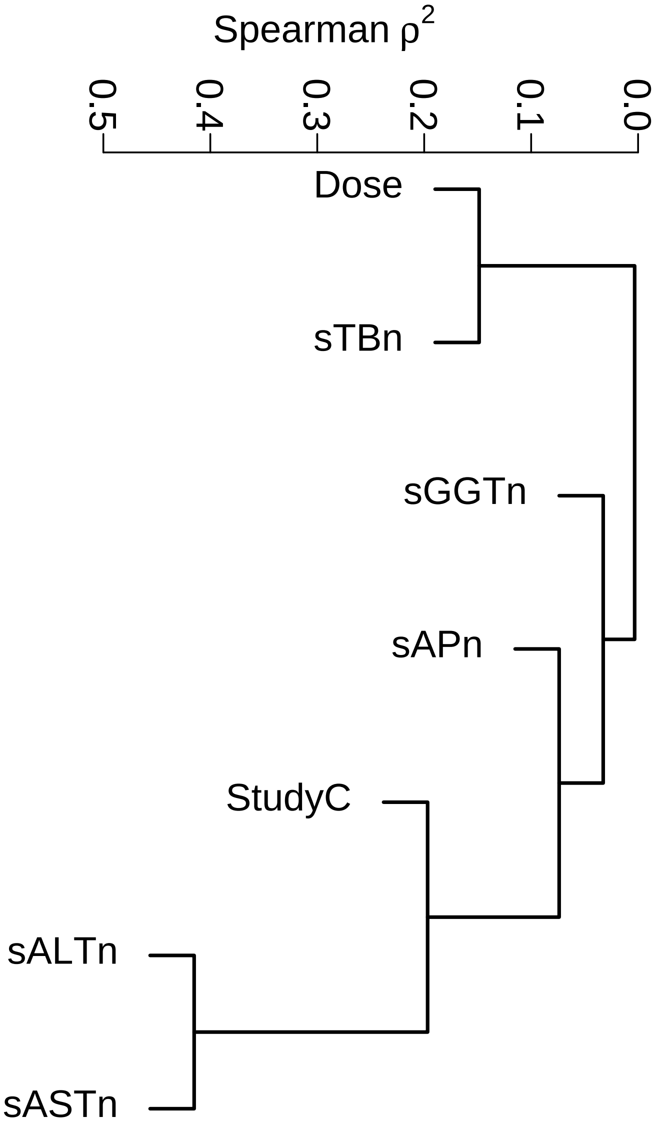

Figure 20: Clustering of slopes, using Spearman ρ^2 as a similarity measure.

Multivariate Analysis

- Proportional odds ordinal logistic model to predict dose from study and 5 slopes
- Initially allowed all interactions with study (global test of interaction: $P = 0.07$)
- Only slope of normalized AP strongly interacted with study
- Spearman ρ for AP slope vs. dose in study C: $P = 0.0004$
- In model with single interaction, only TB and AP were independently affected by increasing dose

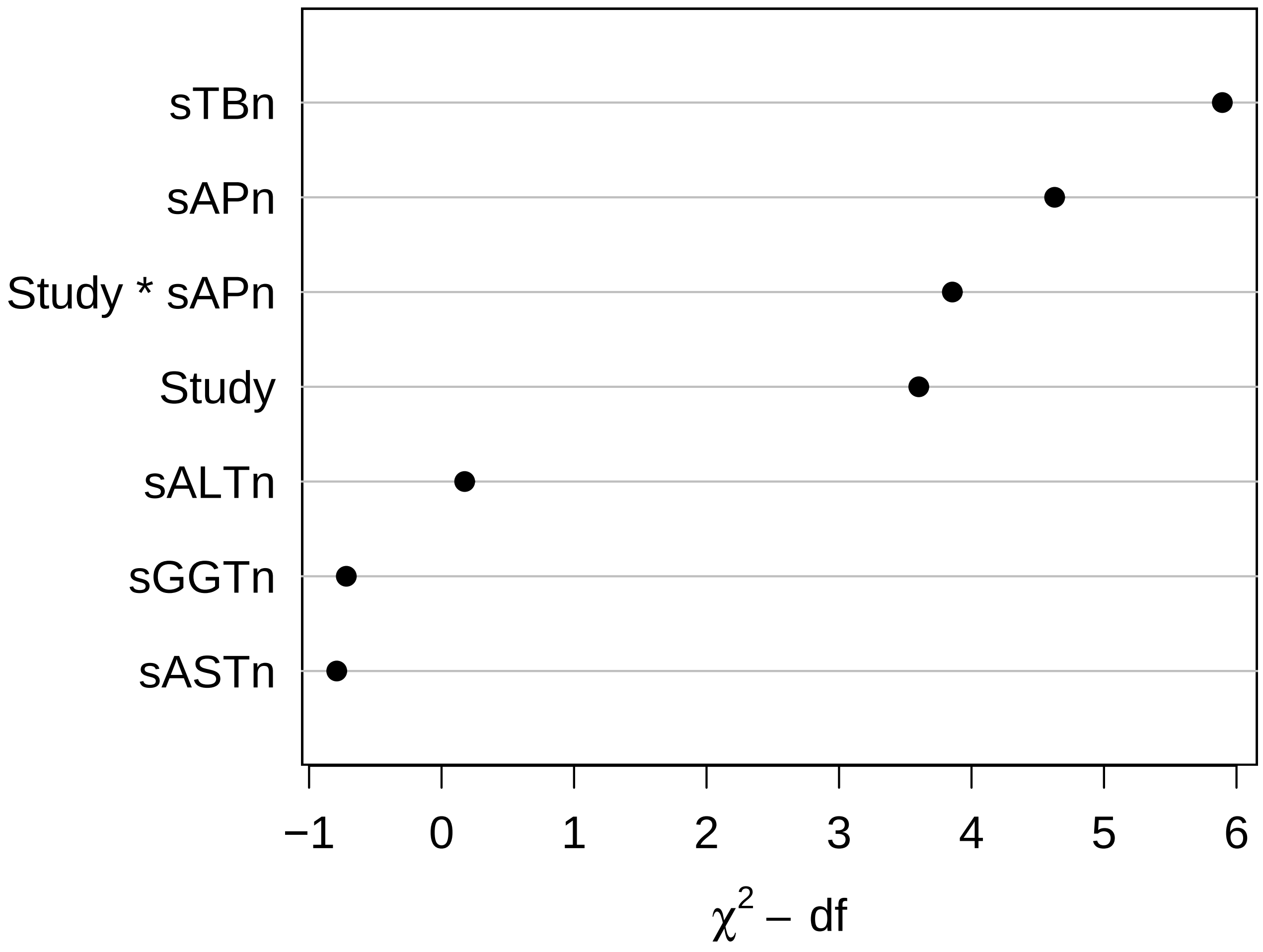

Figure 21: Partial χ^2 statistics penalized for d.f.

Another Study

- Study B (13 subjects)
- No dose effect on slope of bilirubin
- AP slope had strongest correlation with dose

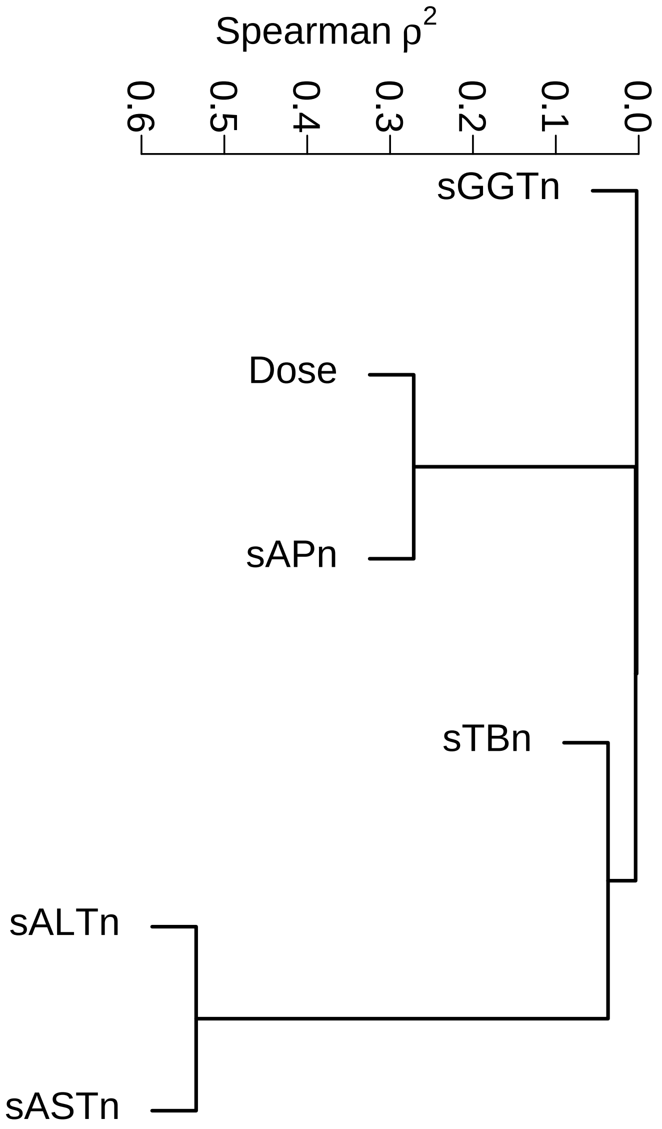

Figure 22: Variable clustering of slopes of normalized clinical chemistry parameters in study B.

Summary

- Graphical exploration of multiple safety response variables has many advantages over generating reams of tables
- Empirical CDFs and extended box plots contain more information than proportion $> k\times$ ULN, mean $\pm$ SD, quartiles
- There are many exploratory analyses to be tapped for safety data
- Some help transform complex multivariate analyses into univariate ones
- Recursive partitioning is a useful exploratory tool
- Exploratory analyses, while not confirming problems or providing causal inference, may provide hypotheses for subject matter experts
- Note: this work was done using only free open-source software: R, LaTeX, Linux

Abstract

It is difficult to design a clinical study to provide sound inferences about safety effects of drugs in addition to providing trustworthy evidence for efficacy. Patient entry criteria and experimental design are targeted at efficacy, and there are too many possible safety endpoints to be able to control type I error while preserving power. Safety analysis tends to be somewhat ad hoc and exploratory. But with the large quantity of safety data acquired during clinical drug testing, safety data are rarely harvested to their fullest potential. Also, decisions are sometimes made that result in analyses that are somewhat arbitrary or that lose statistical efficiency. For example, safety assessments can be too quick to rely on the proportion of patients in each treatment group at each clinic visit who have a lab measurement above two or three times the upper limit of normal.

Safety reports frequently fail to fully explore areas such as

- which types of patients are having AEs?
- what distortions in the tails of the distribution of lab values are taking place?
- which AEs tend to occur in the same patient?
- how to clinical AEs correlate to continuous lab measurements at a given time
- which AEs and lab abnormalities are uniquely related to treatment assigned?
- do preclinically significant measurements at an earlier visit predict AEs at a later visit?

- how can time trends in many variables be digested into an understandable picture?

This talk will demonstrate some of the exploratory statistical and graphical methods that can help answer questions such as the above, using examples based on data from real pharmaceutical trials.

References

[1] E. F. Cook and L. Goldman. Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies. *American Journal of Epidemiology*, 127:626–639, 1988.

[2] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *American Statistician*, 52:181–184, 1998.

[3] P. C. O'Brien. Comparing two samples: Extensions of the t, rank-sum, and log-rank test. *Journal of the American Statistical Association*, 83:52–61, 1988.