Exploratory and Graphical Analysis of Clinical Safety Data

Frank E Harrell Jr and Thomas M Burgan
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat
Slides, S-PLUS and R Code at /presentations

TWENTY-FOURTH ANNUAL
MIDWEST BIOPHARMACEUTICAL STATISTICS WORKSHOP
MUNCIE, INDIANA
21-23 MAY 2001

Outline

1. Tufte on graphical excellence
2. Charts, not tables
3. Extended box plots
4. Clinical trial data
5. Empirical CDFs for lab variables
6. Time trends and clustering of AEs
7. Clustering of lab variables
8. Who is having the most common AE?
9. Which AEs and lab abnormalities are independently related to treatment?
10. Proposal for a default statistical comparison and display of AE incidences (two sample binomial)

Tufte's Views on Graphical Excellence

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space

Tufte, *continued*

- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.”

Charts, Not Tables

- Reviewers find long tables tedious; hard to discern patterns
- Bar charts have many problems
  - hard to show 2-sided CLs
  - bars waste space; hard to show many AEs or categories
- Cleveland's dot charts lead to optimum graphical perception and space usage
- Judicious sorting of categories can aid perception

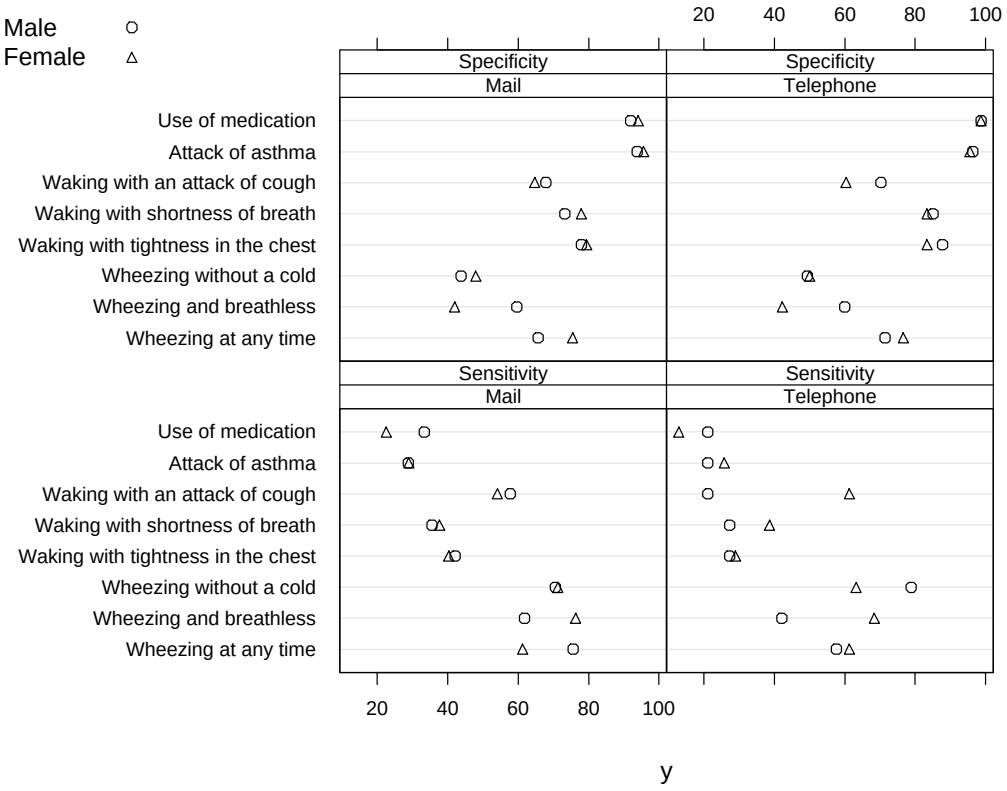

Figure 1: *Example of a dot chart with two superimposed categories per line, stratified by two other variables.*

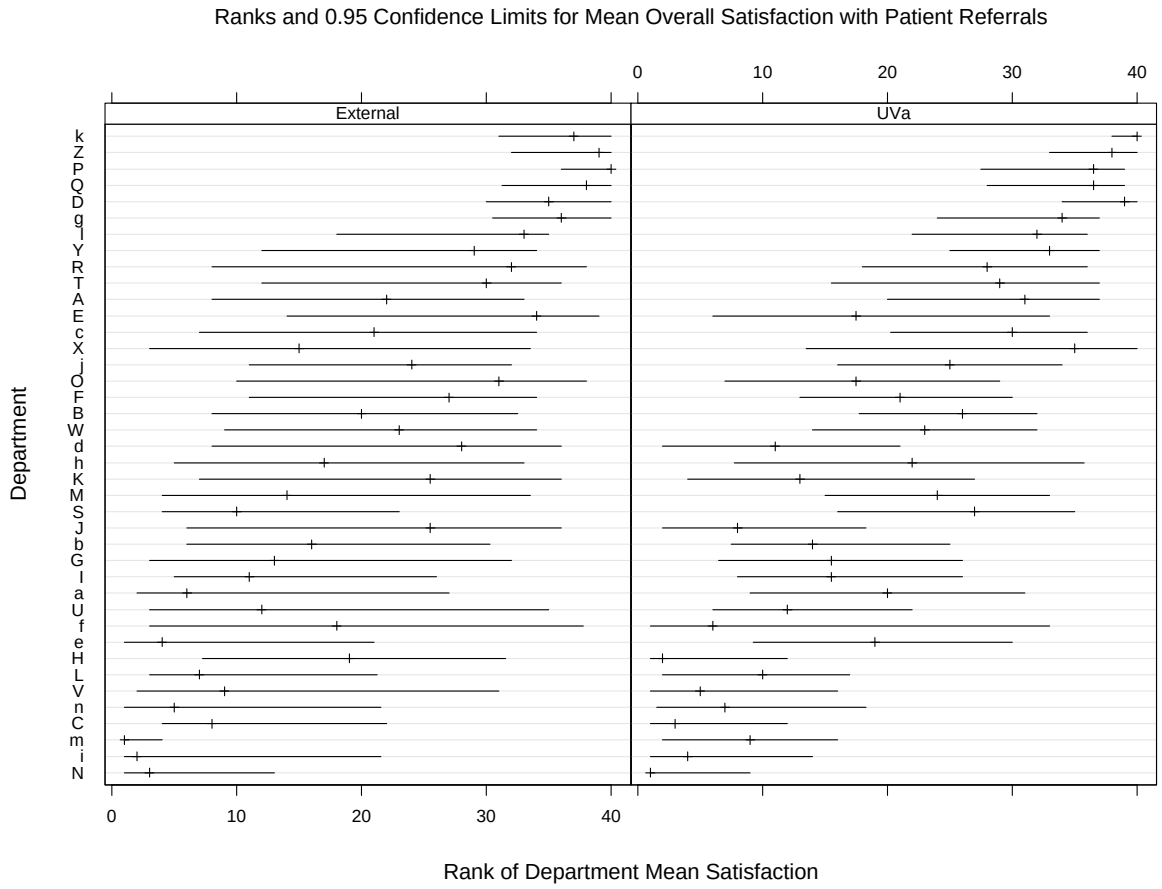

Figure 2: *Dot chart with error bars. Categories are sorted by descending order of the midpoint of the point estimates across the two horizontal groups.*

Extended Box Plots for Lab Data

- Can show not only 3 quartiles but other quantiles
- 0.25, 0.5, 0.75, 0.9 intervals + median and mean

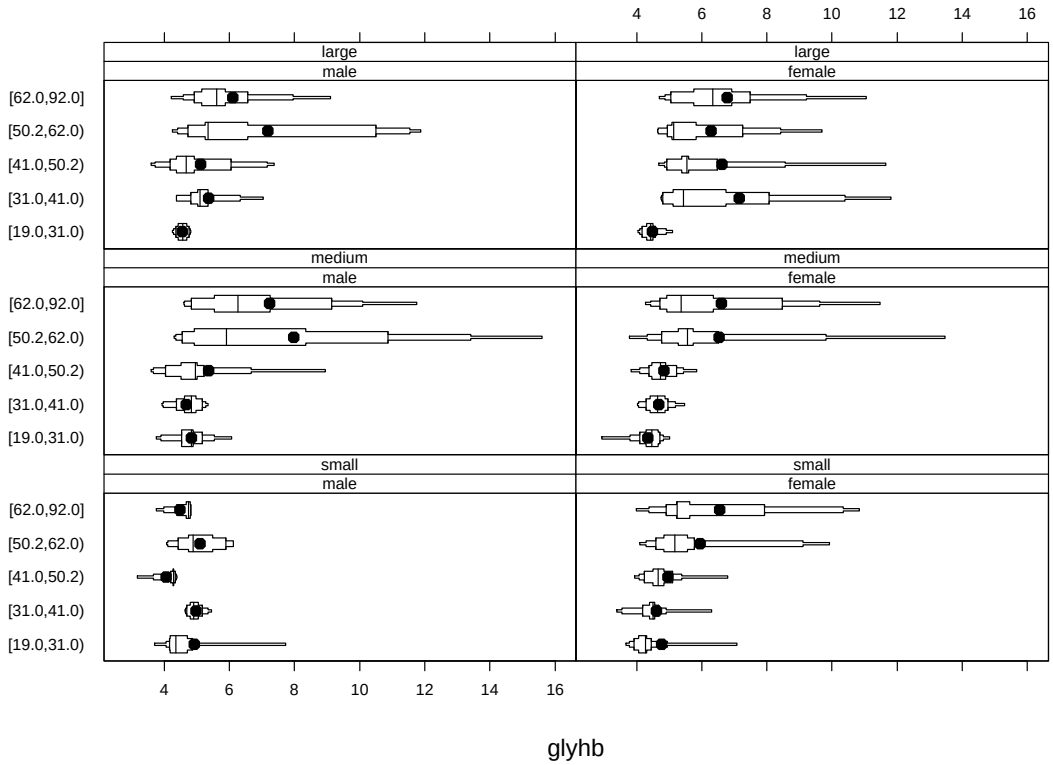

Figure 3: *Extended box plots stratified by two categorical variables (forming panels) and one continuous variable (categorized into quintiles).*

Clinical Trial Data

- A pharmaceutical company generously supplied excellent demographic, AE, clinical lab, 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

- 2 or $3\times$ ULN lose information and are somewhat arbitrary
- 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

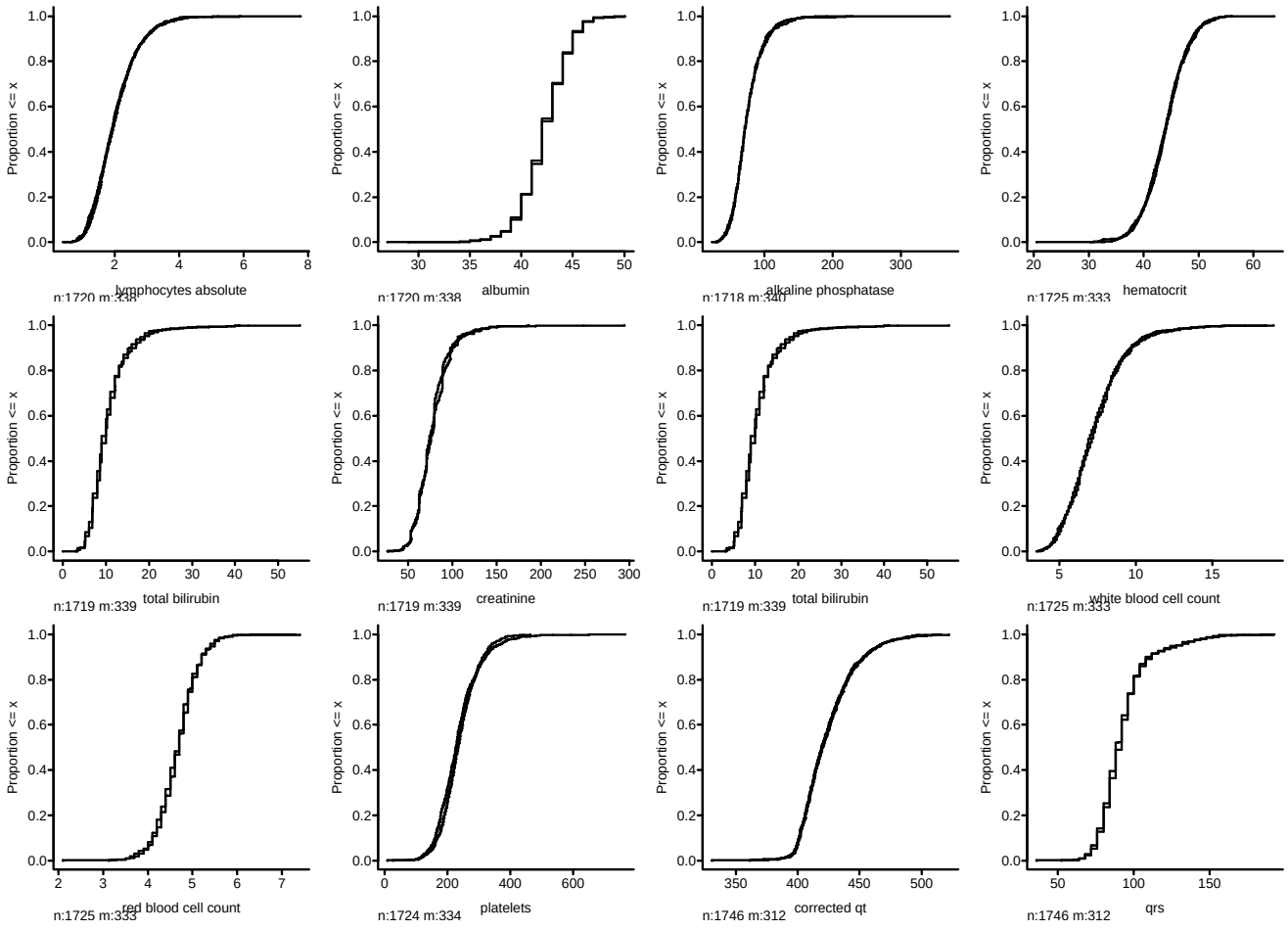

Figure 4: *Empirical CDFs of 12 lab and ECG parameters stratified by treatment group for week 8. CDFs are virtually superimposed.*

Clustering of AEs

- Analyzed 7 AEs having at least 100 episodes
- Which AEs occur together?
- Variable clustering using hierarchical clustering algorithm with similarity matrix = proportion of patients having both AEs (diagonal = 1)

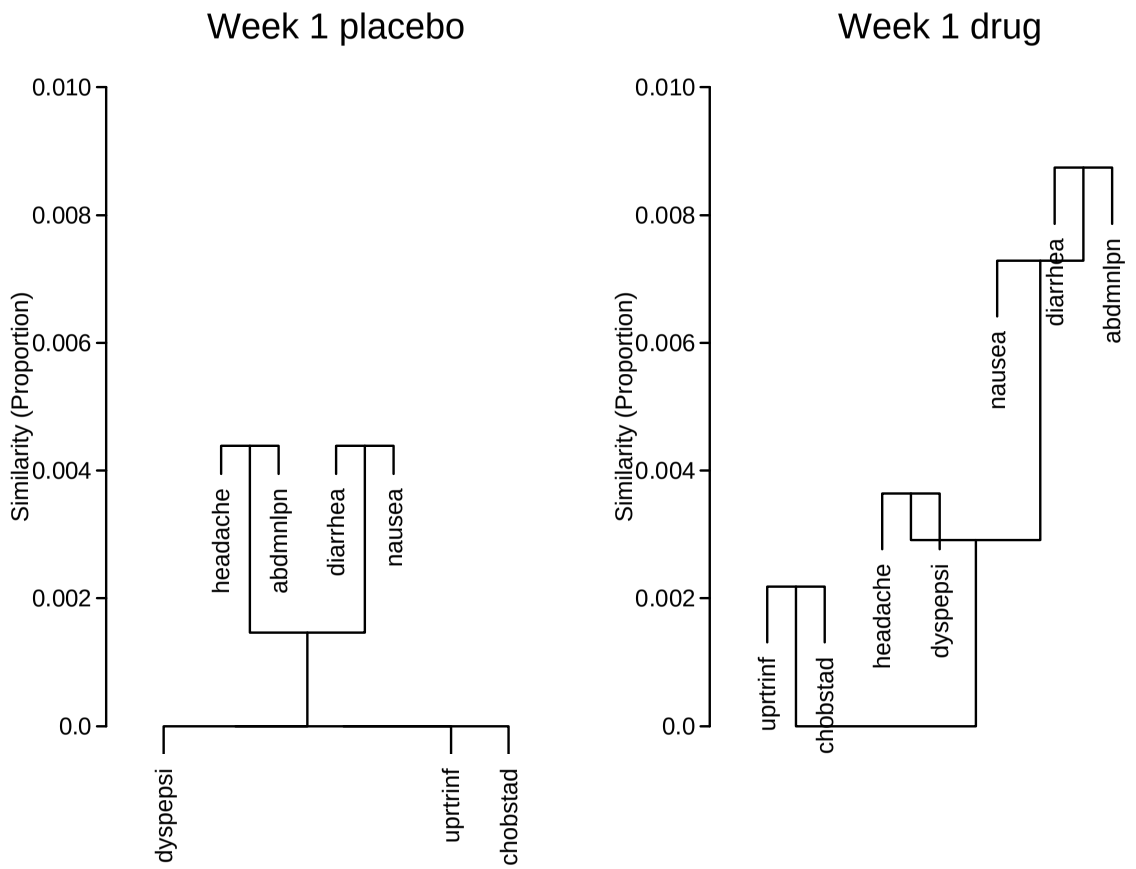

Figure 5: *Variable clustering of AEs at week 1 using proportion of patients having two AEs as similarity measure.*

Time Trends of AE Clustering

- Separately for each AE plot incidence over time for each treatment
- Estimate incidence of pairs of AEs above coincidence levels
- $P_{ij} - P_i P_j$

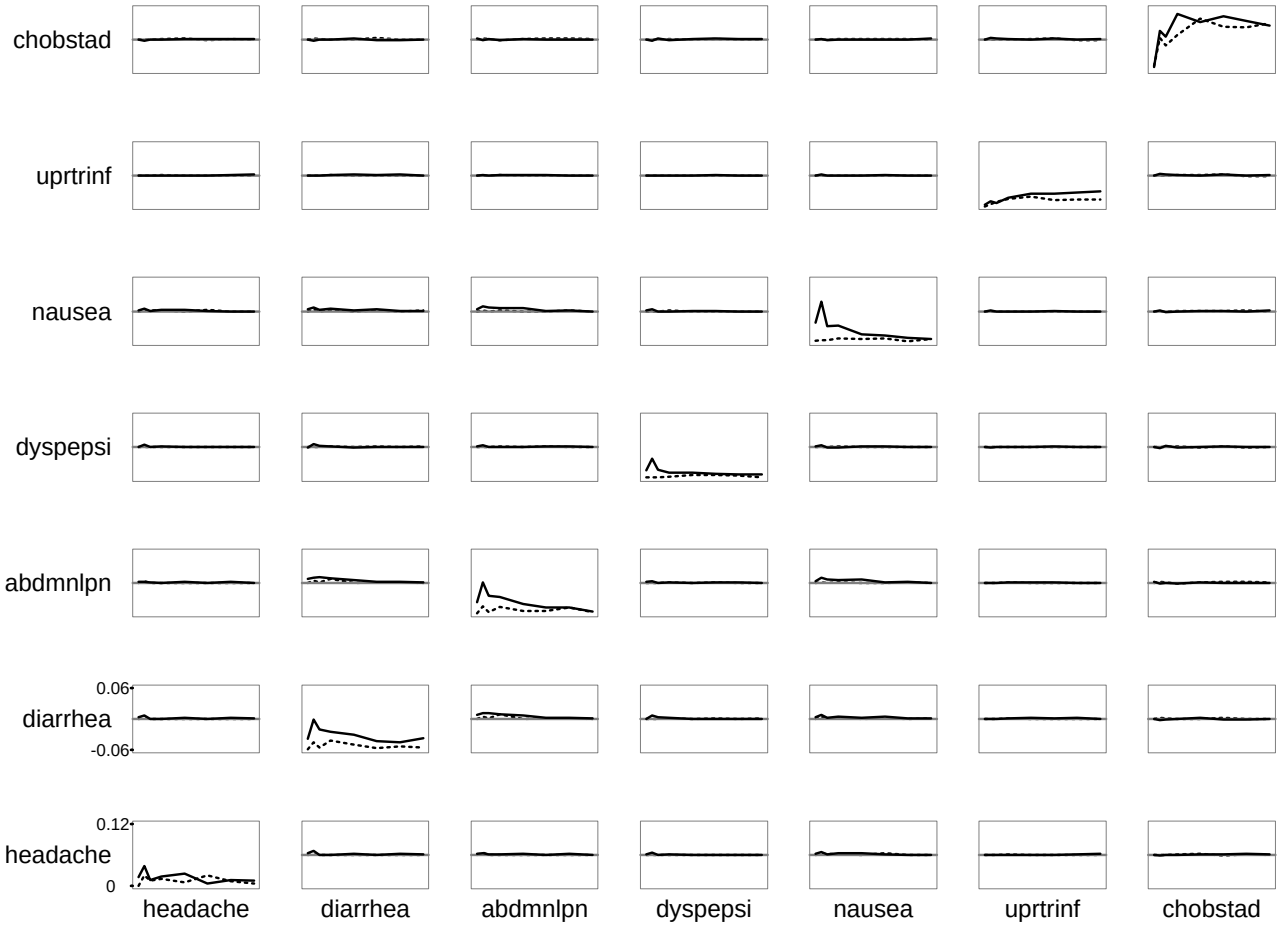

Figure 6: *Time trends in incidence of AEs (diagonal) and chance-corrected joint incidence (off-diagonal). Solid lines represent drug and dotted lines placebo. Horizontal reference lines are at zero (chance level of joint incidence). Week is on x-axes.*

Clustering Lab Variables

- Similarity measure = Spearman ρ^2
- Quantifies strength of monotonic relationships

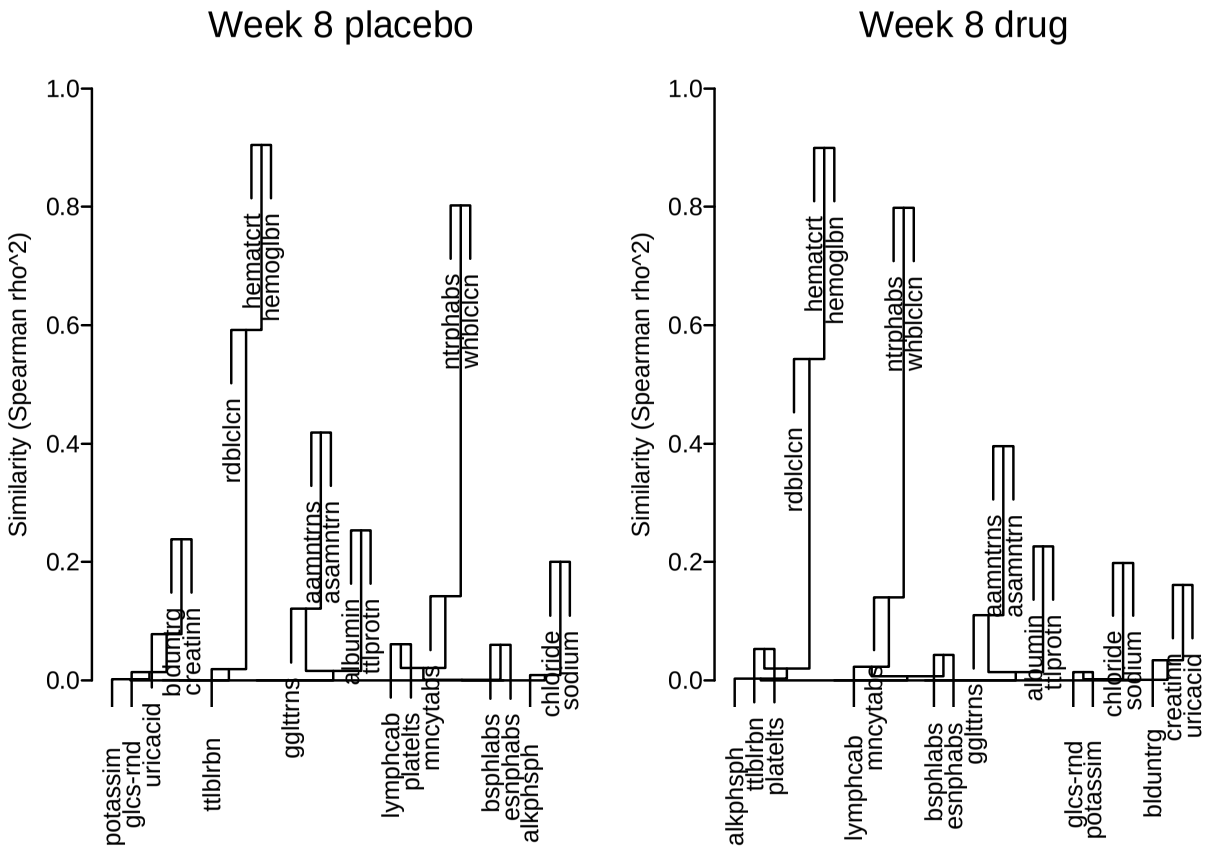

Figure 7: *Variable clustering of lab variables at week 8.*

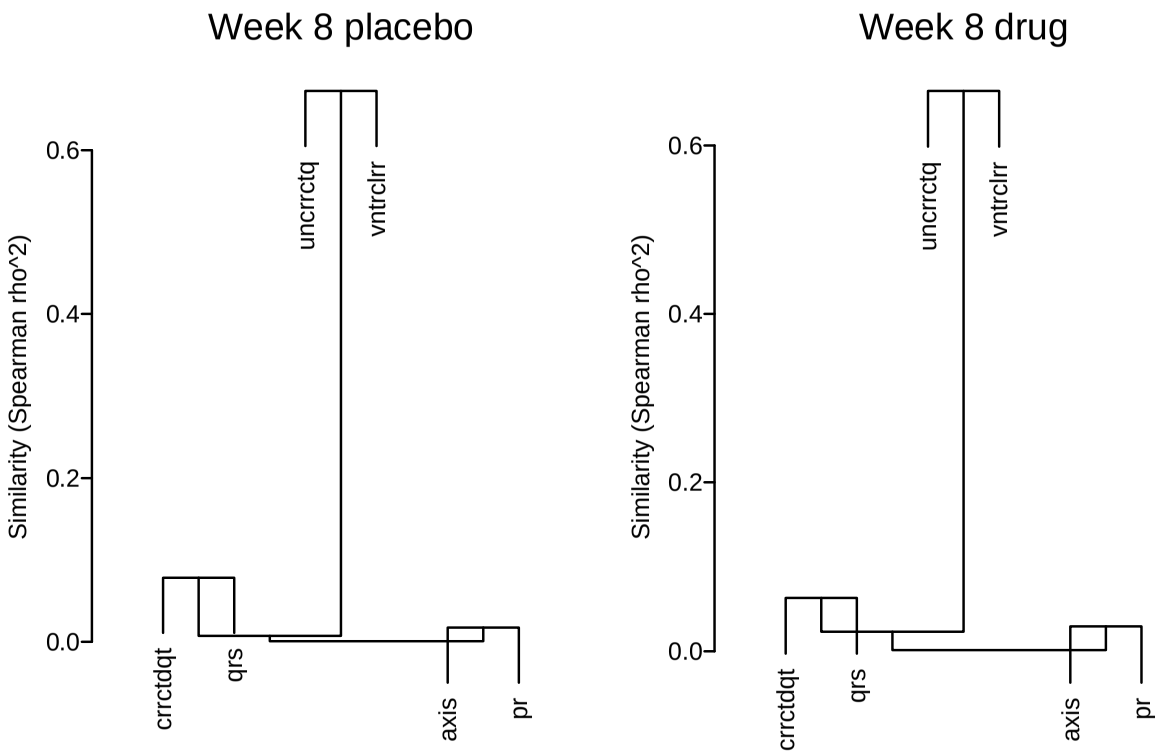

Figure 8: *Variable clustering of ECG parameters at week 8.*

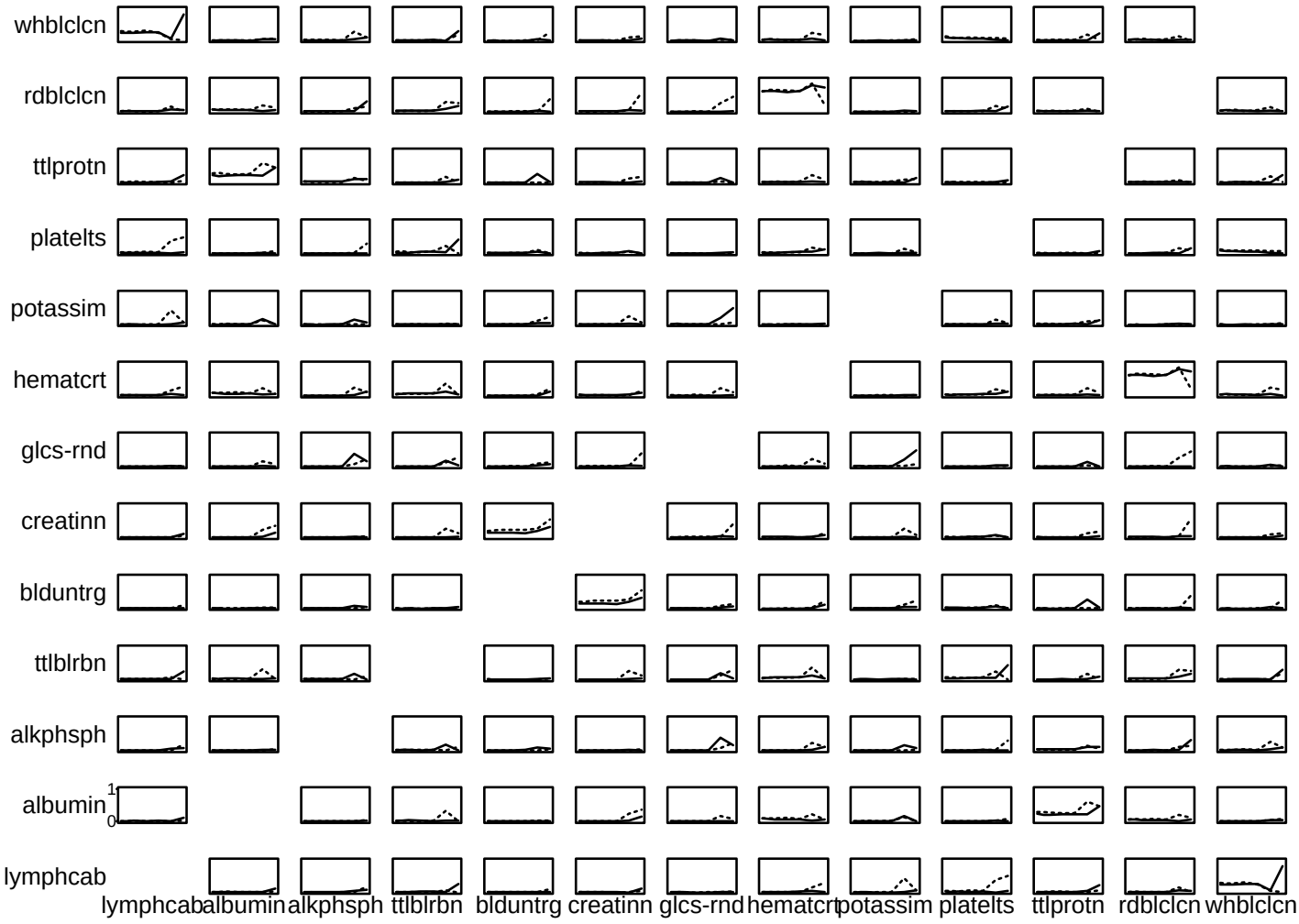

Figure 9: *Time trends in correlation between selected lab variables, stratified by treatment (dotted line = placebo). Y-axes are Spearman ρ^2. Week is on all x-axes.*

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, 13 labs, 8 ECG parameters

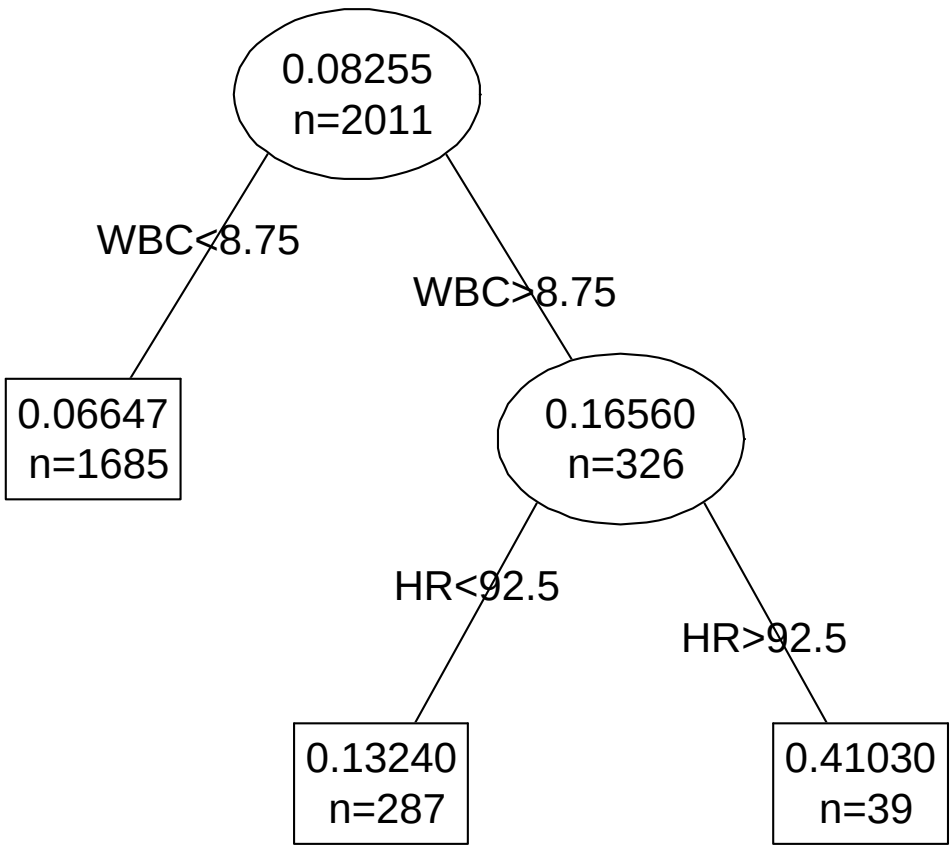

Figure 10: *Regression tree predicting Prob(COAD) at week 8.*

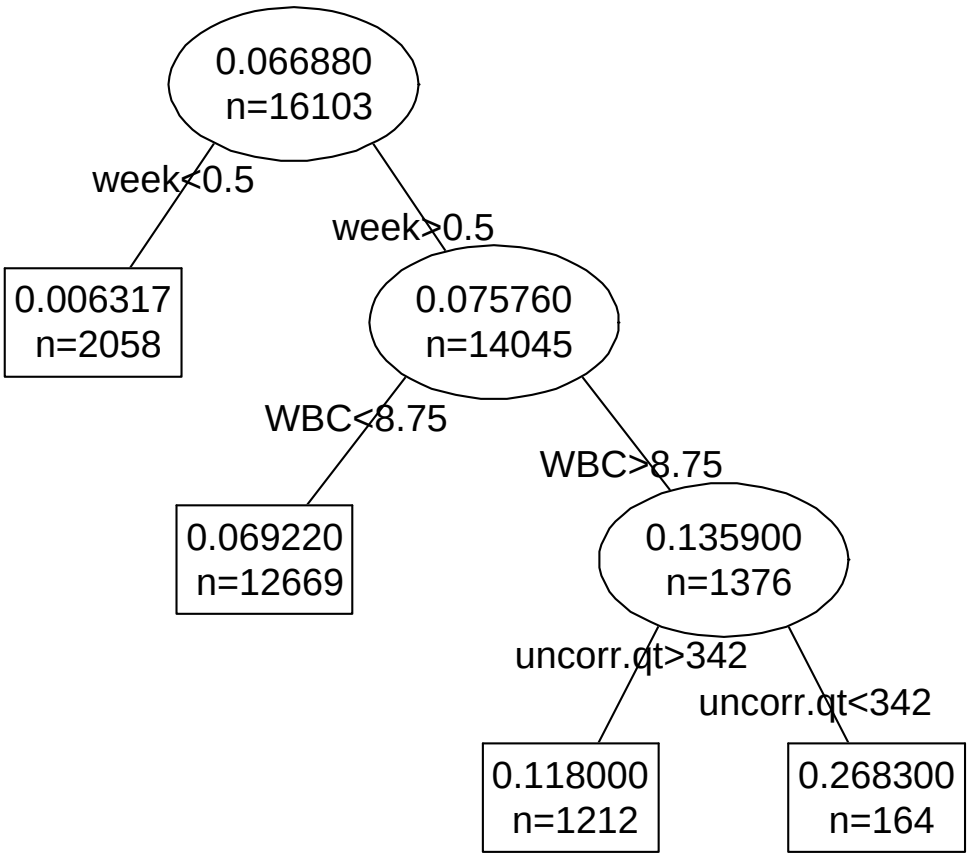

Figure 11: *Regression tree predicting Prob(COAD) at any week.*

Multivariate Analysis of Treatment Differences

- Multiple responses: AEs, labs
- True multivariate methods are cumbersome and make many assumptions
- O'Brien [2] 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

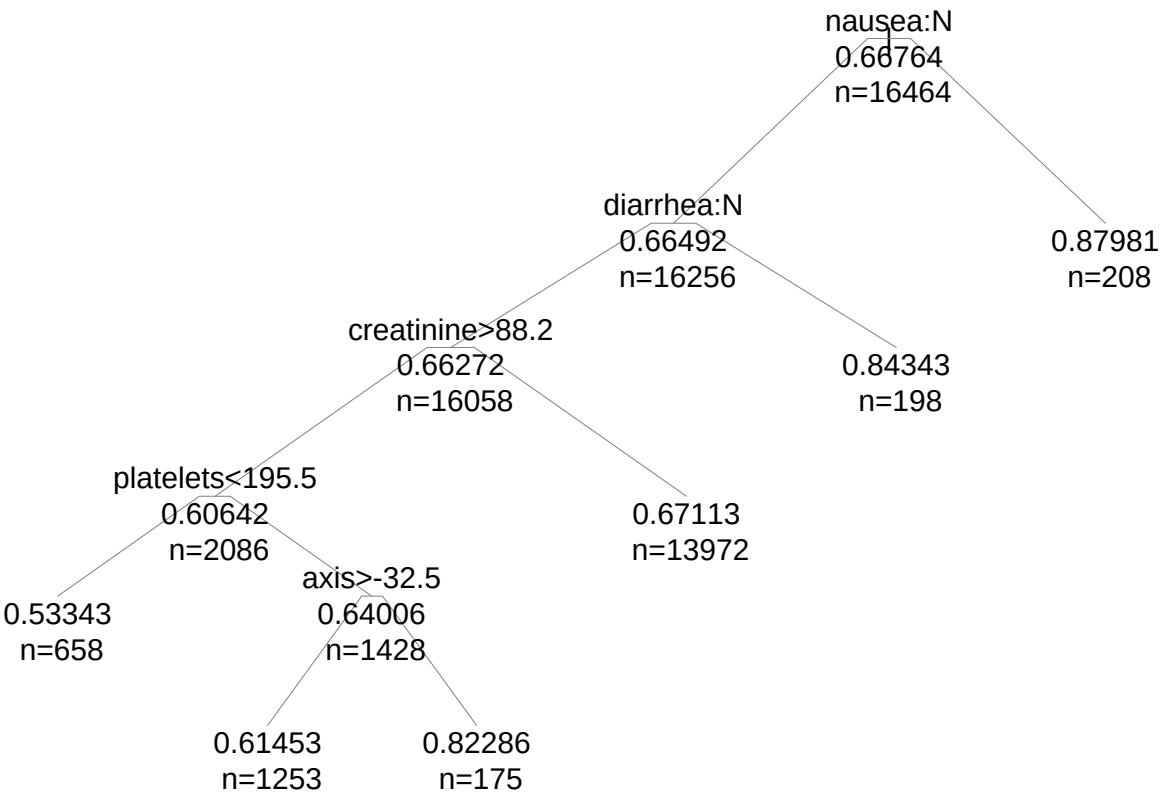

Figure 12: *Regression tree predicting Prob(drug).*

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 of diarrhea without nausea =
  $\frac{1}{2}0.843/.157 = 2.7$

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

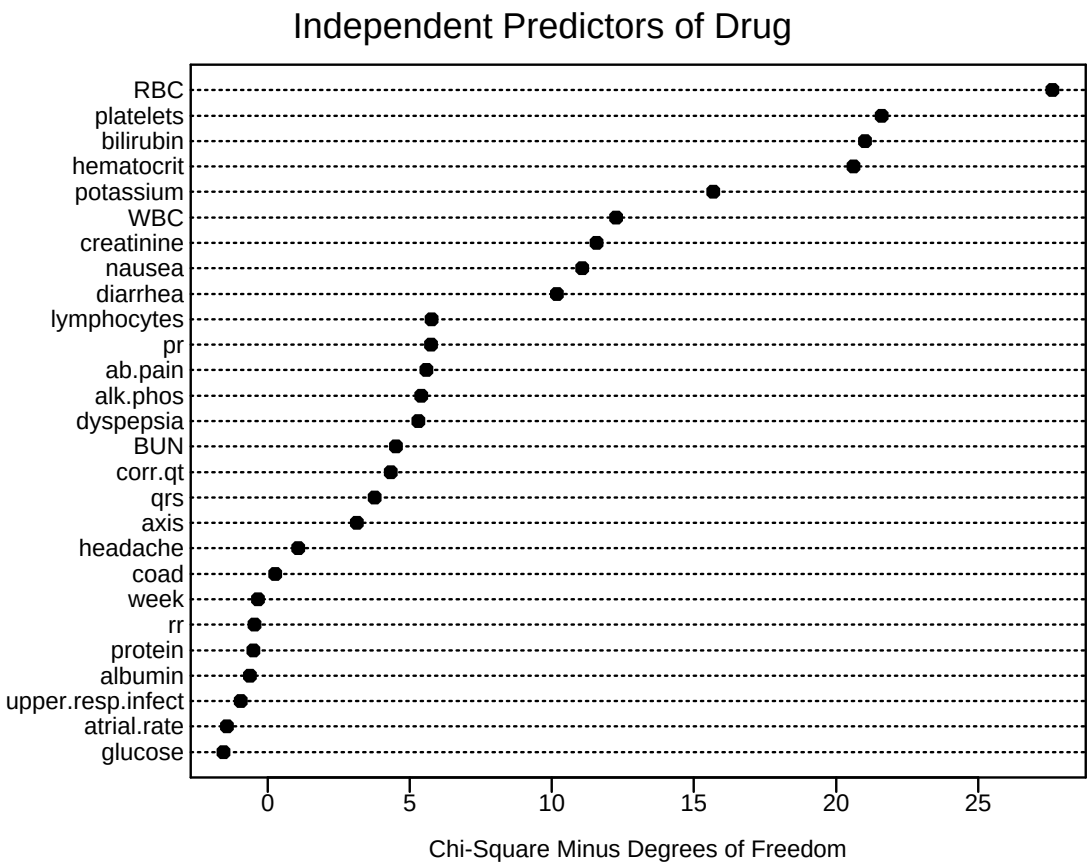

Figure 13: *Degree of partial associations with treatment.*

- Sponsors often omit P-values when comparing AE incidence because study was not powered for safety
- That does not prevent one from computing CLs for risk differences or odds ratios
- Examine least favorable CL for severe AEs to "rule out" significant harm to the patient
- Best summary is Bayesian posterior density of treated:control risk difference and odds ratio
- Can easily compute $P(OR > 1.1 \cup$ risk difference $> 0.025)$ from this

Favored Prior Probabilities

- Gaussian with large variance centered at zero for log odds of AE for control
- Gaussian with mean zero for log odds ratio (treated:control) and variance such that an OR > 4 or $< \frac{1}{4}$ is very unlikely
- Induces a correlation in the prior for P_{drug} and P_{control}
- Uncorrelated β priors would make the computations trivial

Summary

- Charts are preferred to tables even for simple descriptive statistics
- 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
- Need to always present CLs or posterior distributions for possible treatment effects on severe AEs

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 data from a real pharmaceutical trial as a case study.

With the great risk of misinterpretation of P-values (especially large ones), it is alarming that most safety analyses fail to assess statistical evidence for safety concerns because of lack of power. Confidence limits, which are being increasingly relied upon in efficacy analyses, are seldom used in safety reports. More interpretable Bayesian posterior probabilities are used even less frequently. This presentation will provide examples of some graphical reporting formats that could be considered for presenting two-group comparisons of binary safety endpoints when P-values have no meaning.

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] 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.