Measurement, Statistical Consulting, and Computing Issues

Frank E Harrell Jr
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

SmithKline Beecham Biometrics Advisory Board
30 November – 1 December 2000

Outline

1. Measurement issues in chemometrics: What's wrong with ratios?
2. Assay / Microarray measurement issues
3. Problems with ratios in clinical lab data
4. Strategies for analyses that are "correct enough"
5. Applications of pharmacoeconomics to drug discovery
6. Bayesian methods in drug discovery and dose response assessment
7. Problems with discrete survival data
8. Competition from software, and statistical knowledge dissemination
9. Web-based computing as a statistician extender
10. Merck Drug Discovery StatServer
11. SAS vs. S-PLUS

1. Measurement Issues / Ratios

- Unlike differences or log ratios, ratios are asymmetric
- What to subtract from denominator? Most researchers assume zero.
- Ratios have strange distributions
- Kronmal (1993) [2] cited many problems with using ratios in statistical modeling
  - Spurious correlation in using ratio variables even if all component variables of ratios are uncorrelated
  - Division of only the dependent variable by an independent variable can result in regression coefficient estimates for the other independent variables that cause inappropriate conclusions
  - Use of a ratio as an independent variable can result in inadequate adjustment for component variables of the ratio

  - Ratio variables should only be used in a full model containing all the component variables
  - Results of regression analyses incorporating ratios are not readily comparable across studies with different distributions

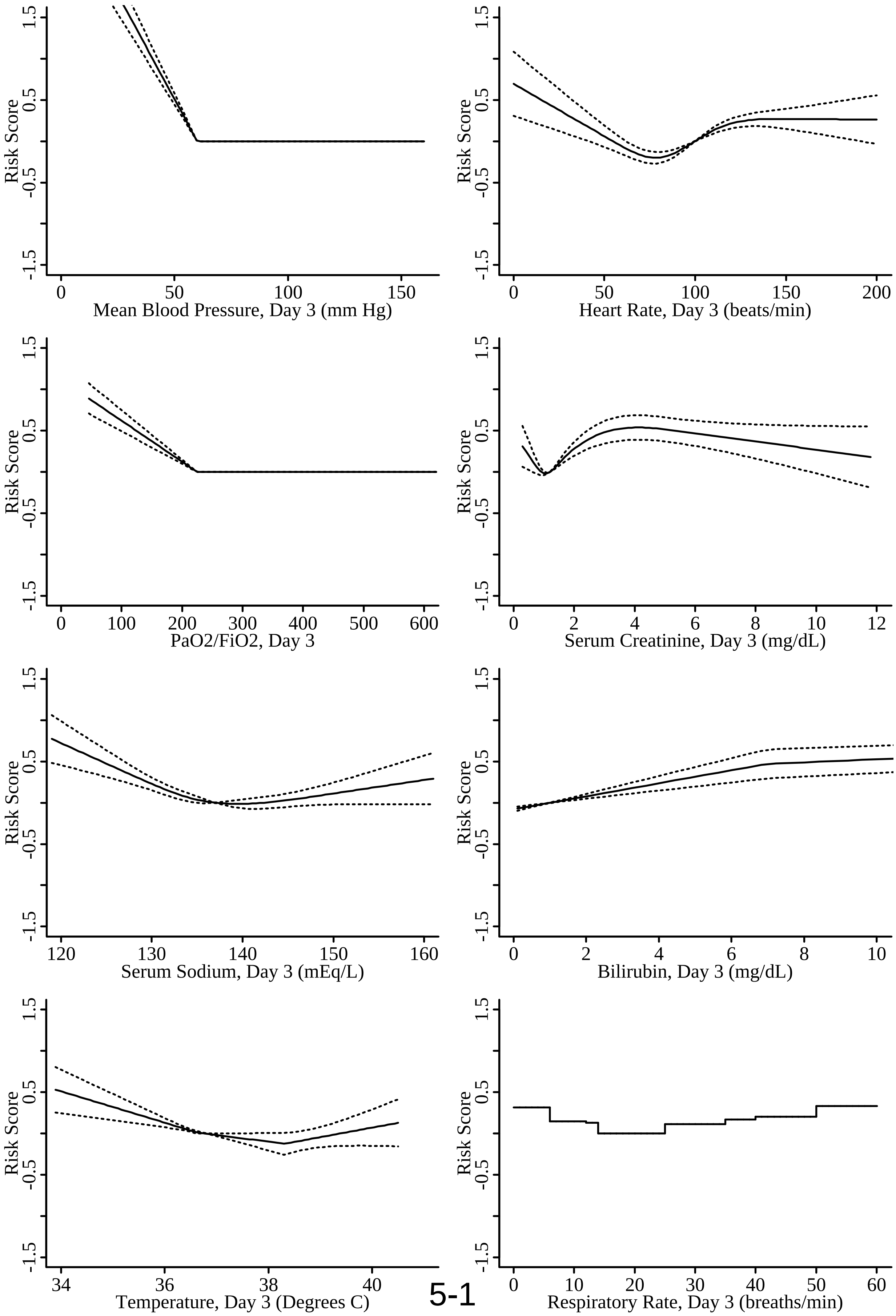
- Kaiser (1989) [1] states that whatever effect measure is chosen (ratio, difference, etc.) should be demonstrated to be uncorrelated with the base value

2. Assay / Microarray Sequencing Issues

- What about values below the lower limit of detectability?
- Not appropriate to eliminate samples
- If using parametric analysis it may not be appropriate to treat values as zero
- Instead, treat them as left-censored
- For rank-based analyses zeros are usually OK
- Wikman *et al.* (2000) [8]: Affymetrix p53 Genechip's 1464 gene chip positions — need to regard "each chip position as a separate entity with its own noise and threshold characteristics"
- Account for row and column effects

3. Problems with Ratios in Clinical Lab Data

- It is common to report the proportion of patients with a lab value $> 3\times$ upper limit of normal
- Problematic: loses information from continuous variables, patients who were “almost abnormal” at baseline will have an easy time moving to the abnormal category
- Problem with non-monotonic risk relationships (e.g., normal range in the middle of the distribution)
- Need to treat lab values as continuous variables without allowing abnormally low values to cancel abnormally high values
- Advantageous to transform to a scale for which “abnormality points” can be added, e.g., log odds, log hazard, log survival time
- Example: scoring physiologic derangements


Risk Score
1.5
0.5
-0.5
-1.5
0
50
100
150
Mean Blood Pressure, Day 3 (mm Hg)
Risk Score
1.5
0.5
-0.5
-1.5
0
50
100
150
200
Heart Rate, Day 3 (beats/min)
Risk Score
1.5
0.5
-0.5
-1.5
0
100
200
300
400
500
600
PaO2/FiO2, Day 3
Risk Score
1.5
0.5
-0.5
-1.5
0
2
4
6
8
10
12
Serum Creatinine, Day 3 (mg/dL)
Risk Score
1.5
0.5
-0.5
-1.5
120
130
140
150
160
Serum Sodium, Day 3 (mEq/L)
Risk Score
1.5
0.5
-0.5
-1.5
0
2
4
6
8
10
Bilirubin, Day 3 (mg/dL)
Risk Score
1.5
0.5
-0.5
-1.5
34
36
38
40
Temperature, Day 3 (Degrees C)
Risk Score
1.5
0.5
-0.5
-1.5
0
10
20
30
40
50
60
Respiratory Rate, Day 3 (breaths/min)

5-1

4. Strategies That are "Correct Enough"

- Example: can one ignore heteroscedasticity in a regression model? It depends.
- Rather than doing a weighted analysis, make the transformation of variables an integral part of the analysis
- Regression splines for independent variables
- Nonparametric smoothers in transform-both-sides generalized additive models, e.g. AVAS (Tibshirani [6]): transform Y to make variance of residuals independent of $\hat{Y}$.
- When only two variables are being analyzed at a time and only a P-value is needed, use rank test and rank correlation
- Use robust rank-based regression (Cox, proportional odds model) for multiple variables

5. Pharmacoeconomics in Drug Discovery

- Can’t think of any applications except for Bayesian decision analysis incorporating costs (below)
- Are plenty of applications to drug *development*
- See Senn (1996) [5] for probabilistic decision analysis for portfolio management of compounds

6. Bayesian Methods in Drug Discovery and Dose Response Assessment

- In drug discovery type I error is of concern
- Bayesian prior distributions are usually a better way to deal with multiplicity
- Can incorporate prior distribution for the chances that a biomarker is an efficacy marker or for probability of monotonicity of dose-response
- Can do formal Bayesian decision analysis that incorporates costs of false positives and false negatives
- Bayesian methods have small-sample exactness without conditioning on only part of the data
- Opportunity for statisticians to be called on more by other reseachers: “It takes time to be a Bayesian”

7. Discrete Survival Data

- May be best to use a method that is dedicated to heavily tied data, e.g. Prentice-Gloeckner [4]
- Investigate using Efron likelihood in Cox model
- Try randomly breaking ties and using ordinary Cox likelihood

8. Competition from Software / Statistical Knowledge Dissemination

- Not enough statisticians to go around
- Researchers are using statistical software and choosing the wrong software (e.g., Excel)
- Newsletter on choice of software
- Ongoing short course series (e.g., Statistical Thinking in Biomedical Research) emphasizing study design, bias, measurement, precision, power, graphics, demos (“what a statistician does with data”)
- Clients should know almost as much about statistics as we know about biology

9. Web-Based Computing as a Statistician Extender

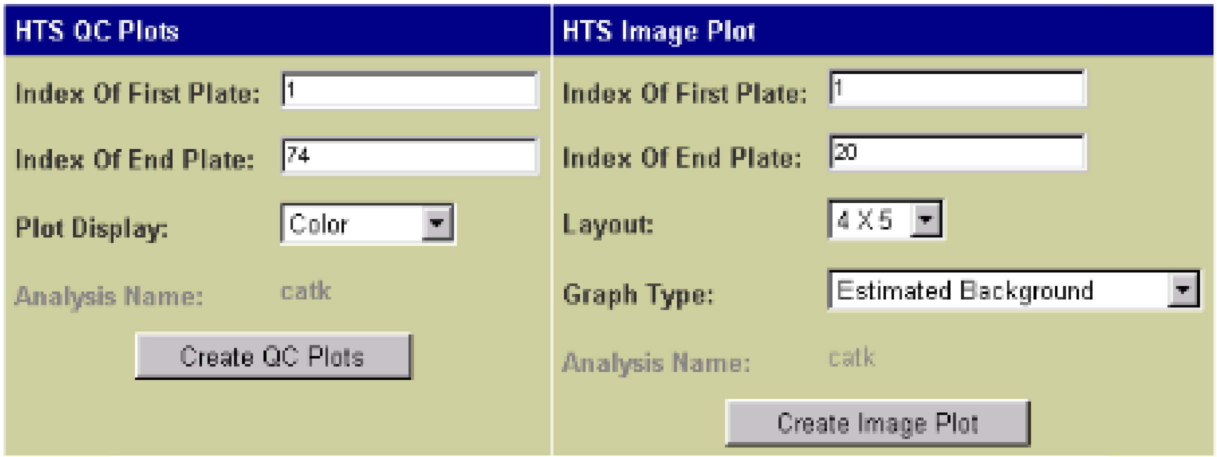
- “Safe Statistics”: Pikounis, Gunter, Liaw, Pajni (2000) [3]
- Statistical strategies that
  - “Produce useful answers ‘most’ of the time
  - Indicate where answers may not be useful
  - Have ‘adequate’ performance
  - Handle missing values and other data problems
  - Are tuned to user skill level

- In practice this means
  - Graphics
  - Well designed user interface
  - Resistant methods
  - Fewest assumptions possible (nonparametric procedures)
  - Use of subject matter knowledge whenever possible”
- Statisticians can control which methods are distributed or emphasized to non-statisticians

10. Merck S-PLUS HTS StatServer

- S-PLUS StatServer is web based, no special client software
- 96-3456 well plates for HTS assays in drug discovery
- Take into account positional effects within plates (esp. edge effects), changing background response and assay sensitivity, trends, cycles, shifts, missing values
- Used by 3000 scientists with little formal stat knowledge, fewer than 10 statisticians to support them
- Drilling down after potential problems seen (e.g., analysis by rows or by columns)
- Heavy on graphics and nonparametric trends
- Error messages to users are also E-mailed to statisticians

- Detailed usage accounting data



HTS QC Plots

Index Of First Plate: 1

Index Of End Plate: 74

Plot Display: Color

Analysis Name: catk

HTS Image Plot

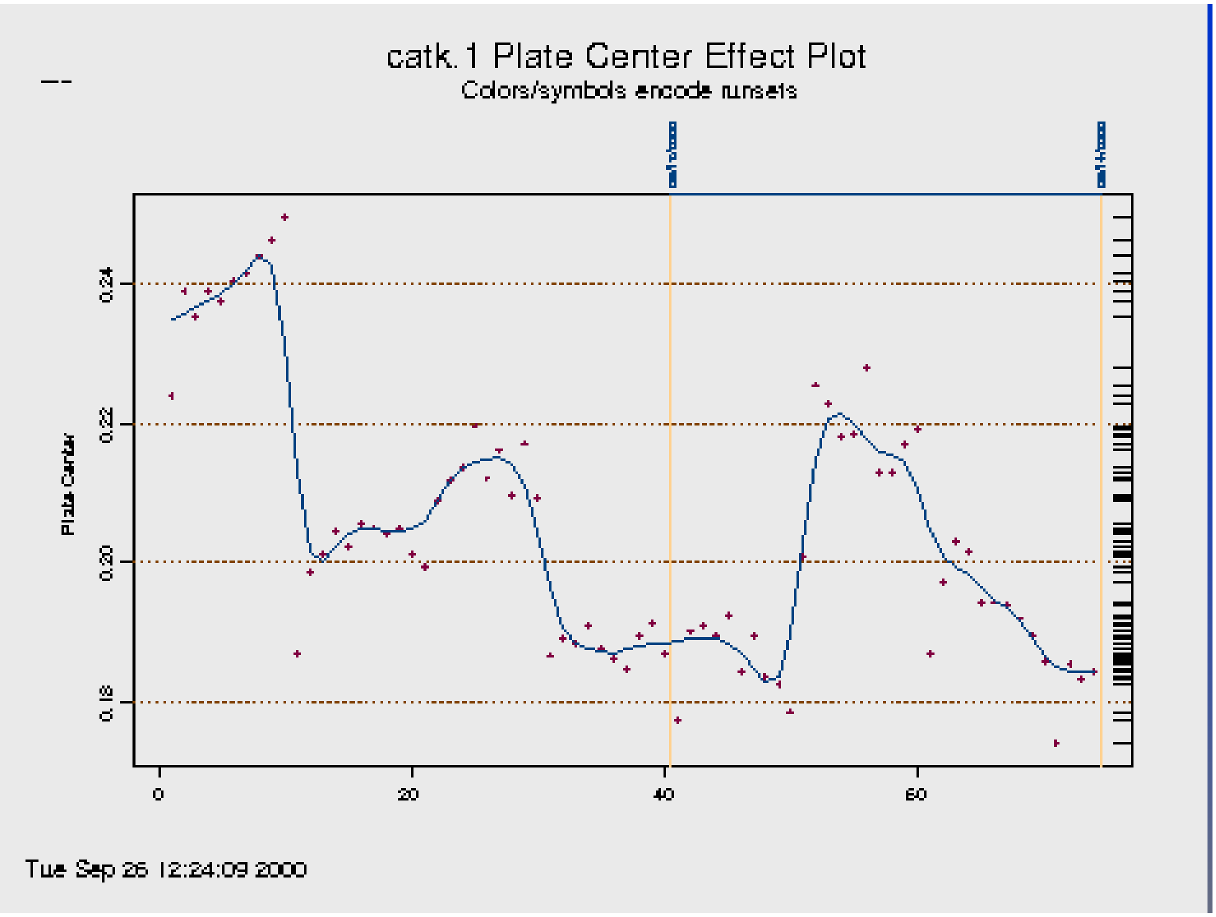
Index Of First Plate: 1

Index Of End Plate: 20

Layout: 4 X 5

Graph Type: Estimated Background

Analysis Name: catk



11. S-PLUS and SAS

- SAS 8 has narrowed the gap to 5 years
- Major advantages of S-PLUS:

1. No distinction between DATA and PROC steps
2. No macro language; all commands are “live”.

   Example:

   ```
   if(is.category(x) |
   is.character(x) |
   length(unique(x)) < 20)
   table(x) else quantile(x)
   ```

3. Many more data types than SAS, users can add their own attributes to data (e.g., flag strange or imputed values)
4. Truly interactive
5. Graphics
6. S-PLUS 2000 comes with 2900 functions

7. Language is extendible and relatively simple to program; user-written functions are written in the same language used by the developers; statisticians world-wide are writing functions
8. Speed of implementation of modern methods (StatLib)
9. Methods for modeling, exploratory data analysis, missing data, graphics after model fitting, bootstrap, table making, much more [7]

10. No need for output delivery system:

   - All entities are objects, allowing all functions to communicate directly
   - Special methods for formatting output, e.g.:

```
# create LaTeX table (alt: HTML)
latex(summary(marker ∼ age+sex))
# logistic regression model with
# regression splines, interactions
f ← lrm(y ∼ rcs(age,5)*sex +
         rcs(pressure,4))
f              # ordinary printout
plot(f)        # show fitted shapes
Function(f)    # create S+ function
sascode(Function(f))    # SAS code

# typeset fit in algebraic form
w ← latex(f)
html(w)        # convert to HTML
# future: convert to XML with
# embedded MathML
xml(f)
```

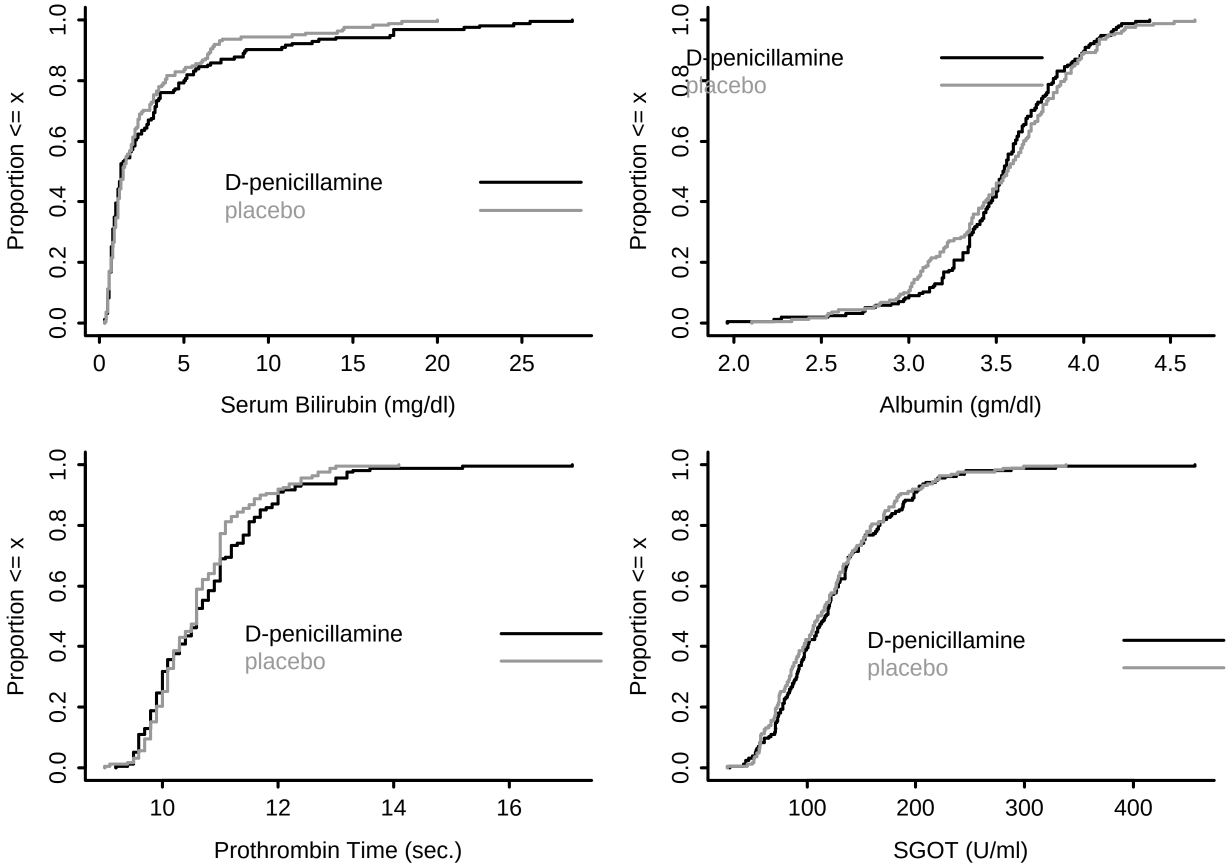
Table Making Example

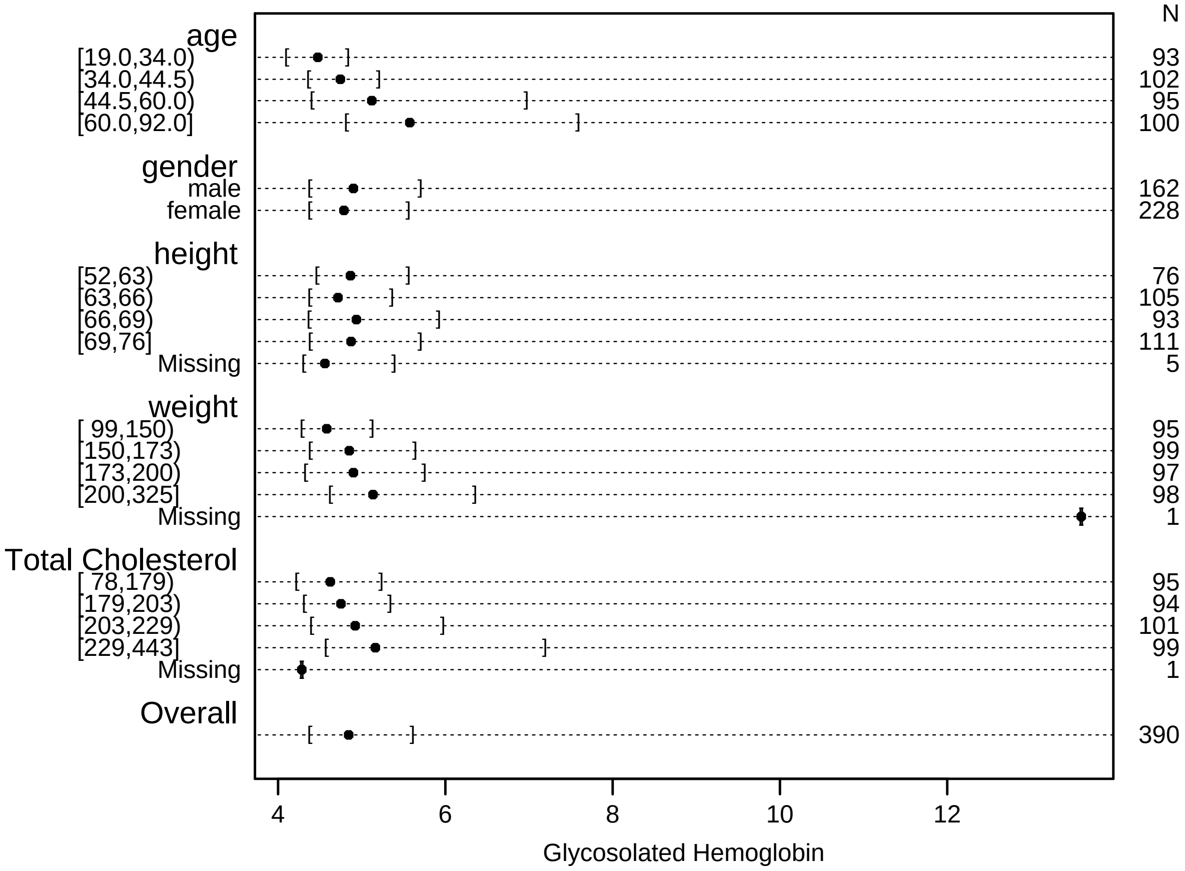
```
s ← summary(drug ∼ bili + albumin +
               stage + protime + sex + age +
               spiders,  method='reverse')
latex(s, npct='both')
```

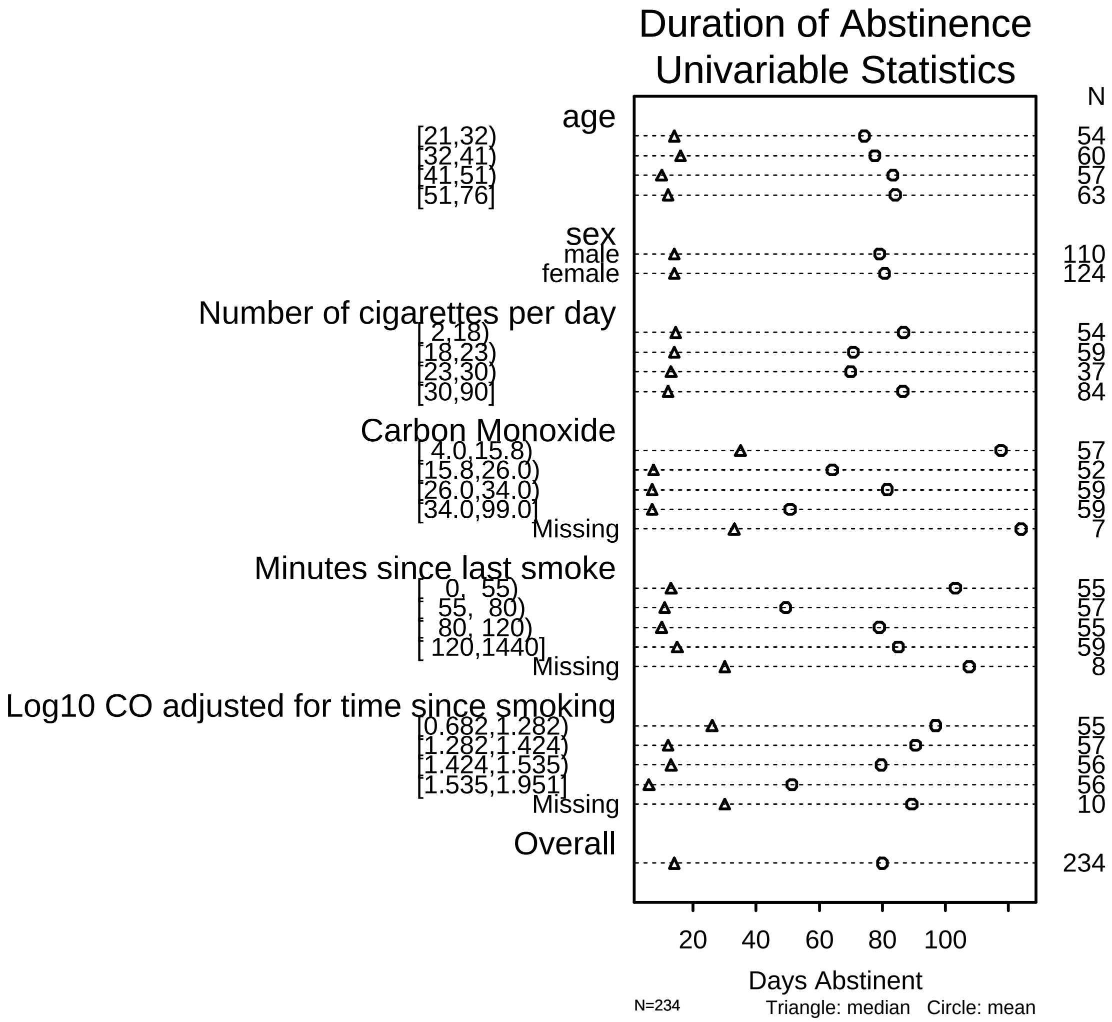
	N	D-penicillamine $(N = 154)$	placebo $(N = 158)$
Serum Bilirubin (mg/dl)	418	0.725 **1.300** 3.600	0.800 **1.400** 3.200
Albumin (gm/dl)	418	3.34 **3.54** 3.78	3.21 **3.56** 3.83
Histologic Stage, Ludwig Criteria : 1	412	**3%** $\frac{4}{154}$	**8%** $\frac{12}{158}$
2		**21%** $\frac{32}{154}$	**22%** $\frac{35}{158}$
3		**42%** $\frac{64}{154}$	**35%** $\frac{56}{158}$
4		**35%** $\frac{54}{154}$	**35%** $\frac{55}{158}$
Prothrombin Time (sec.)	416	10.0 **10.6** 11.4	10.0 **10.6** 11.0
Sex : female	418	**90%** $\frac{139}{154}$	**87%** $\frac{137}{158}$
Age	418	41.4 **48.1** 55.8	43.0 **51.9** 58.9
Spiders	312	**29%** $\frac{45}{154}$	**28%** $\frac{45}{158}$

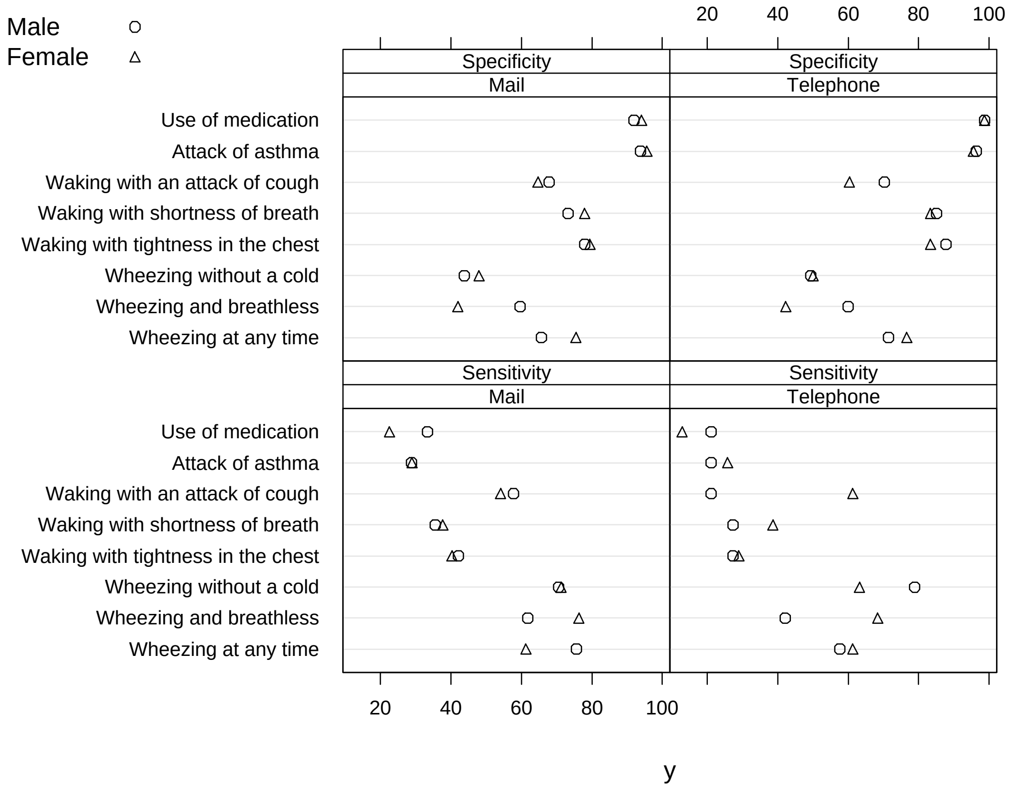
a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.

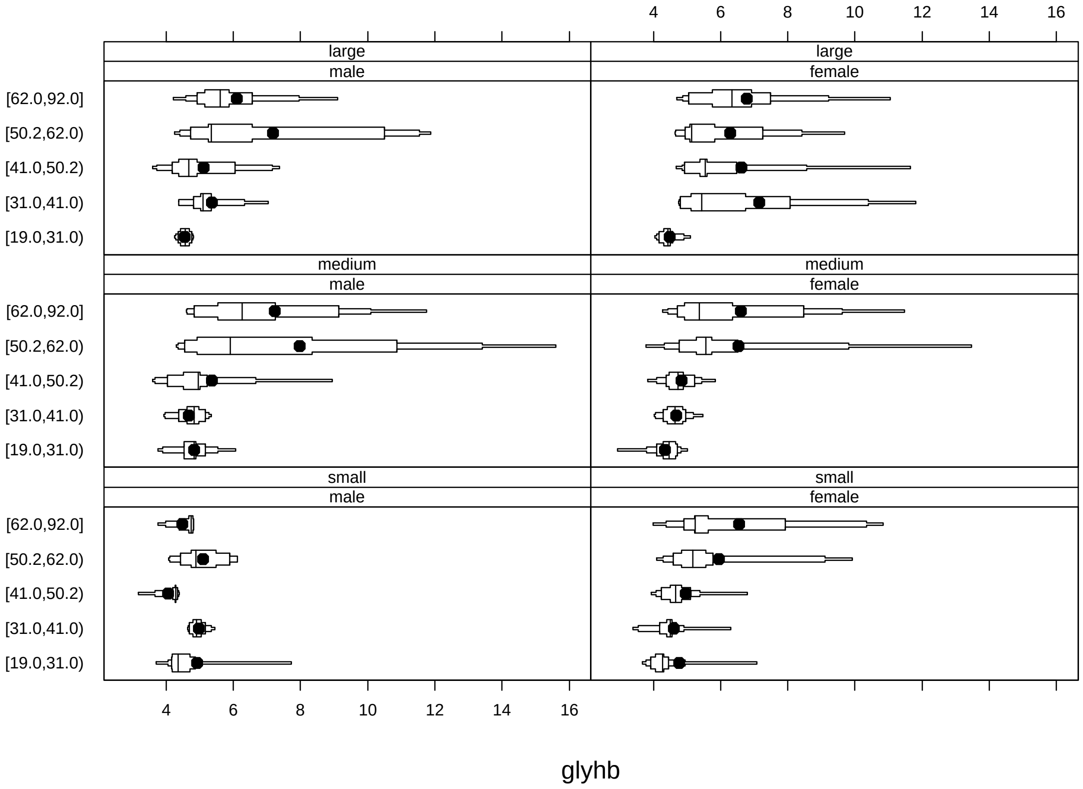
N is the number of non–missing values.

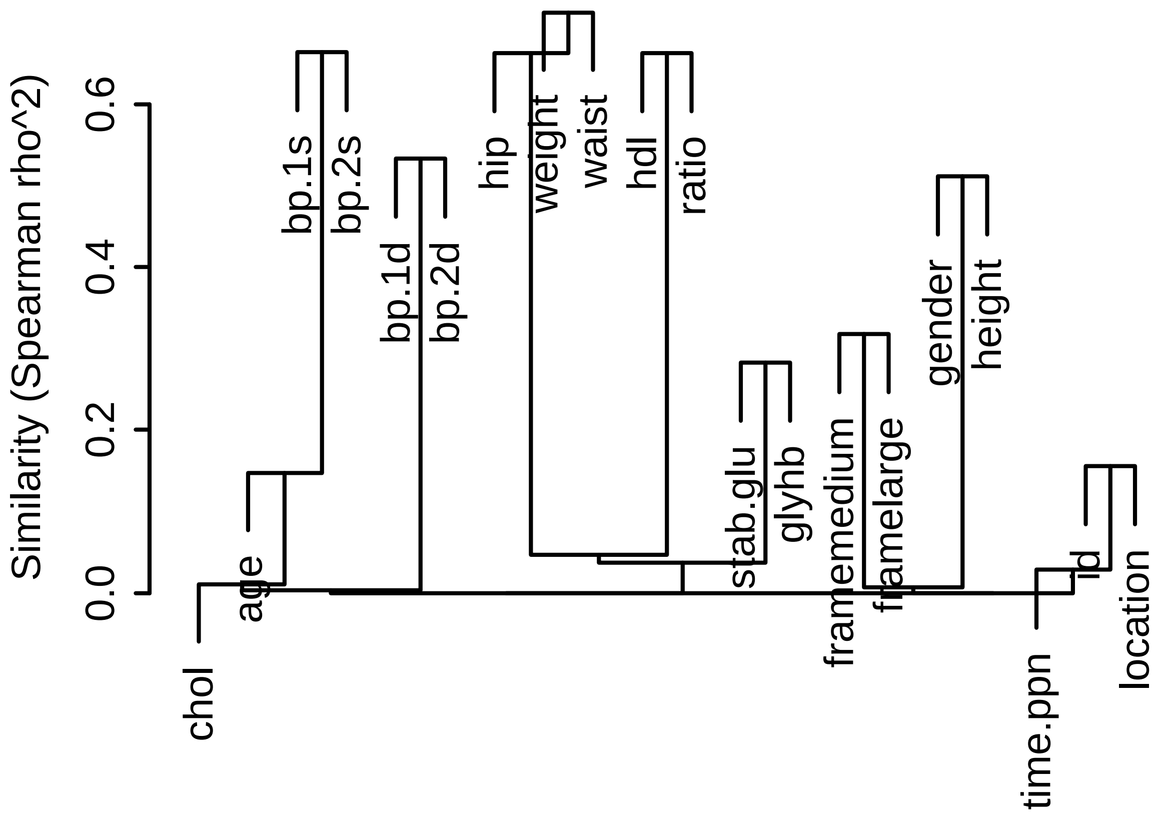

Proportion <= x
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Serum Bilirubin (mg/dl)
D-penicillamine
placebo
Proportion <= x
2.0
2.5
3.0
3.5
4.0
4.5
Albumin (gm/dl)
D-penicillamine
placebo
Proportion <= x
10
12
14
16
Prothrombin Time (sec.)
D-penicillamine
placebo
Proportion <= x
100
200
300
400
SGOT (U/ml)
D-penicillamine
placebo

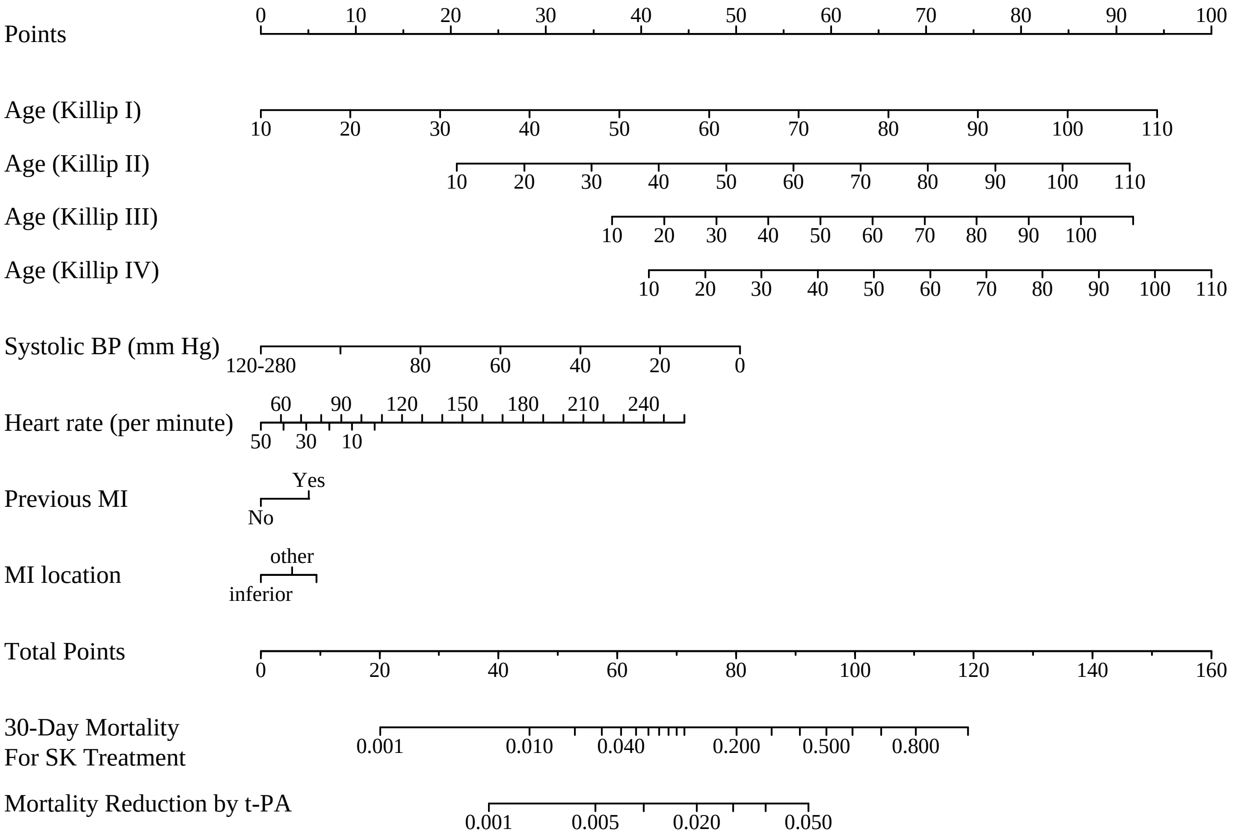

N
age
[19.0,34.0)
[34.0,44.5)
[44.5,60.0)
[60.0,92.0]
93
102
95
100
gender
male
female
162
228
height
[52,63)
[63,66)
[66,69)
[69,76]
Missing
76
105
93
111
5
weight
[99,150)
[150,173)
[173,200)
[200,325]
Missing
95
99
97
98
1
Total Cholesterol
[78,179)
[179,203)
[203,229)
[229,443]
Missing
95
94
101
99
1
Overall
390
4
6
8
10
12
Glycosolated Hemoglobin


Duration of Abstinence
Univariable Statistics
N
age
[21,32) 54
[32,41) 60
[41,51) 57
[51,76] 63
sex
male 110
female 124
Number of cigarettes per day
[2,18) 54
[18,23) 59
[23,30) 37
[30,90] 84
Carbon Monoxide
[4.0,15.8) 57
[15.8,26.0) 52
[26.0,34.0) 59
[34.0,99.0] 59
Missing 7
Minutes since last smoke
[0, 55) 55
[55, 80) 57
[80, 120) 55
[120,1440] 59
Missing 8
Log10 CO adjusted for time since smoking
[0.682,1.282) 55
[1.282,1.424) 57
[1.424,1.535) 56
[1.535,1.951] 56
Missing 10
Overall 234
20
40
60
80
100
Days Abstinent
N=234
Triangle: median Circle: mean


Male ○
Female △
Specificity
Mail
Specificity
Telephone
Sensitivity
Mail
Sensitivity
Telephone
Use of medication
Attack of asthma
Waking with an attack of cough
Waking with shortness of breath
Waking with tightness in the chest
Wheezing without a cold
Wheezing and breathless
Wheezing at any time
Use of medication
Attack of asthma
Waking with an attack of cough
Waking with shortness of breath
Waking with tightness in the chest
Wheezing without a cold
Wheezing and breathless
Wheezing at any time
20
40
60
80
100
20
40
60
80
100
y


4
6
8
10
12
14
16
large
male
large
female
[62.0,92.0]
[50.2,62.0)
[41.0,50.2)
[31.0,41.0)
[19.0,31.0)
medium
male
medium
female
small
male
small
female
glyhb


Similarity (Spearman rho^2)
0.0
0.2
0.4
0.6
chol
age
bp.1s
bp.2s
bp.1d
bp.2d
hip
weight
waist
hdl
ratio
stab.glu
glyhb
framemedium
framelarge
gender
height
time.ppn
id
location


Points
0 10 20 30 40 50 60 70 80 90 100
Age (Killip I)
10 20 30 40 50 60 70 80 90 100 110
Age (Killip II)
10 20 30 40 50 60 70 80 90 100 110
Age (Killip III)
10 20 30 40 50 60 70 80 90 100
Age (Killip IV)
10 20 30 40 50 60 70 80 90 100 110
Systolic BP (mm Hg)
120-280 80 60 40 20 0
Heart rate (per minute)
60 90 120 150 180 210 240
50 30 10
Previous MI
Yes
No
MI location
other
inferior
Total Points
0 20 40 60 80 100 120 140 160
30-Day Mortality
For SK Treatment
0.001 0.010 0.040 0.200 0.500 0.800
Mortality Reduction by t-PA
0.001 0.005 0.020 0.050

pbc
19 Variables 418 Observations

bili : Serum Bilirubin (mg/dl)

```
   n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 418 0       98      3.221  0.50  0.60  0.80  1.40  3.40  8.03 14.00

lowest :  0.3  0.4  0.5  0.6  0.7, highest: 21.6 22.5 24.5 25.5 28.0
```

albumin : Albumin (gm/dl)

```
   n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 418 0       154     3.497 2.750 2.967 3.243 3.530 3.770 4.010 4.141

lowest : 1.96 2.10 2.23 2.27 2.31, highest: 4.30 4.38 4.40 4.52 4.64
```

stage : Histologic Stage, Ludwig Criteria

```
   n missing unique  Mean
 412 6       4      3.024

1 (21, 5%), 2 (92, 22%), 3 (155, 38%), 4 (144, 35%)
```

protime : Prothrombin Time (sec.)

```
   n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 416 2       48      10.73  9.60  9.80 10.00 10.60 11.10 12.00 12.45

lowest :  9.0  9.1  9.2  9.3  9.4, highest: 13.8 14.1 15.2 17.1 18.0
```

sex : Sex

```
   n missing unique
 418       0      2

male (44, 11%), female (374, 89%)
```

fu.days : Time to Death or Liver Transplantation

```
   n missing unique Mean    .05    .10    .25    .50    .75    .90    .95
 418 0       399    1918  245.1  606.8 1092.8 1730.0 2613.5 3524.2 4040.6

lowest :   41   43   51   71   77, highest: 4500 4509 4523 4556 4795
```

age : Age

```
   n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 418 0       345     50.74 33.84 36.37 42.83 51.00 58.24 64.30 67.92

lowest : 26.28 28.88 29.56 30.28 30.57
highest: 74.52 75.00 75.01 76.71 78.44
```

spiders : Spiders

```
   n missing unique
 312     106      2

absent (222, 71%), present (90, 29%)
```

References

[1] Lee Kaiser. Adjusting for baseline: Change or percentage change? *Statistics in Medicine*, 8:1183–1190, 1989.

[2] R. A. Kronmal. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society A*, 156:379–392, 1993.

[3] Bill Pikounis, Bert Gunter, Andy Liaw, and Neeraj Pajni. Automated analysis software for screening using S-PLUS StatServer. S-PLUS Users Conference, October 2000.

[4] R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with applications to breast cancer data. *Biometrics*, 34:57–67, 1978.

[5] Stephen Senn. Some statistical issues in project prioritization in the pharmaceutical industry. *Statistics in Medicine*, 15:2689–2702, 1996.

[6] Robert Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83:394–405, 1988.

[7] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999.

[8] Friedrik P. Wilman, Ming-Lan Lu, Thomas Thykjaer, Sanne H. Olesen, Lars D. Andersen, Carlos Cordon-Cardo, and Torben F. Orntoft. Evaluation of the performance of a p53 sequencing microarray chip using 140 previously sequenced bladder tumor samples. *Clinical Chemistry*, 46:1555–1561, 2000.