Cox regression analysis

- The most common analysis seen in medical journals
- Uses a regression algorithm to determine how well a finding “A” explains results “B”
- It allows use of data the way it is available in the real world by adjusting for:
  - Different periods of follow-up for subjects
  - Allows a range of ages and both sexes to be included accounting for the impact of those variables
  - Accounts for other differences between subjects that may impact outcome such as smoking, stage, grade or laboratory test results
Why Cox is so valuable

• Generating a multivariate analysis.
• By using this multivariate approach (in contrast to the univariate Kaplan Meyer or life table, etc.)
  – You can keep all the (limited) data together in the analysis rather than splitting by age, sex or duration
  – You don’t have too few outcomes (the bane of any medical researcher’s existence)
  – You can account of the impact of each variable on the outcome

How does Cox do it?

• It uses regression algorithms to assign a particular hazard ratio (mortality ratio) to each variable across the population studied to best explain the overall outcome in question
• It is done with software programs such as SPSS or SAS or Strata which are purchased and loaded with the data and then the Cox button is pushed after options (or defaults) chosen
### Outcome - hazard ratio

- A hazard ratio (such as 1.45) reflecting the relative risk of an outcome based on:
  - one integer to the next averaged across all values (GGT 100 vs 101 or by year of age) or
  - between “categories” such as male and female or
  - between bands of values against a reference band (GGT <65, 65-100, 101+) or
  - (as commonly done) between % grouping such as tertiles or quintiles.

### Outcome - survival curve

- Much like a Kaplan Meyer plot but adjusts for impact of other variables (age and smoking in the example)
- Could also be shown as a hazard curve instead
Confidence intervals

• Along with the hazard ratio, you should see confidence intervals which mostly replace “p values”
  – Typically expressed as the hazard ratios at the upper and lower end of the 95% CI meaning there is only a 5% chance the real risk would lie outside the interval
  – Width of the 95% CI - largely dependent on # of outcomes (deaths). 30 or more per cell is good.
  – If the CIs overlap, then the hazard is not demonstrated to be different at a 95% significance level.
    • May not actually be different or
    • May have too few deaths and too wide a CI to tell

Limitations of Cox
Cox is designed to sort risk between independent variables

- Fructosamine and HbA1c will have some mortality risk allocated to each if both are included yet both are measuring the same thing- glycosylation caused by elevated BS and only one should be included.
- Less extreme examples are LFTs which are collinear (highly correlated with one another) for outcomes such as death
- Although collinearity can be evaluated mathematically, understanding the biology is more useful

Missing important variables

- The distribution of cholesterol values vary by age and sex with the lowest values in young females who will also have the lowest mortality
- This leads to conclusions about cause and effect that may be erroneous (Applicants tested 1999-2006 with VS in 2010)

<table>
<thead>
<tr>
<th></th>
<th>Haz Ratio</th>
<th>95% CI for Haz Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOLESTEROL</td>
<td>1.005</td>
<td>1.004</td>
</tr>
<tr>
<td>CHOLESTEROL</td>
<td>1.001</td>
<td>1.000</td>
</tr>
<tr>
<td>SEX</td>
<td>.671</td>
<td>.665</td>
</tr>
<tr>
<td>AGE</td>
<td>1.098</td>
<td>1.098</td>
</tr>
</tbody>
</table>
Impact on group vs. subgroup

GGT (& age-sex) by Cox

Handling “U” or “J” shaped mortality

- Transform data so it works in Cox
  - Create synthetic variables that are more linear by “binning”
  - Use “splines” separated at knots
  - All starts with understanding your data
HbA1c study - an example of missing variables, confidence intervals & -

- HbA1c values <5%, 5-5.4%, 5.5-5.9%, 6-6.4%, etc.
- Had the following hazard ratios (MRs):

<table>
<thead>
<tr>
<th></th>
<th>&lt; 5%</th>
<th>5-5.4%</th>
<th>5.5-5.9%</th>
<th>6-6.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>1</td>
<td>1.25</td>
<td>1.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Women</td>
<td>1</td>
<td>1.02</td>
<td>1.3</td>
<td>1.6</td>
</tr>
</tbody>
</table>

HbA1c study contd.
Missing variables, confidence intervals & -

- Wide overlapping CIs (based on 15-88 deaths/cell)
  - plausible but poor support
- Included only age as covariate with sex split in their main result table
  - differences in smoking (8.4 – 10 - 16 - 19.4%) and CV risk factors included only in a secondary analysis
ROC

A “simple” (simplistic) graphical tool to evaluate how well any test result predicts the presence (sensitivity) and absence (specificity) of a condition or the risk of an event

• Generates a graph and an AUC (area under the curve) identical to the “c statistic”.
  – No association has AUC of 0.5; perfect association is 1.0 (or 0.0 if the association is negative)

• A univariate analysis requiring care that the association is not really based on other, not-included, variables such as age and sex

• Similar limitation to Cox regarding U or J shaped data

ROC - Example

Mortality for men age 60-69
ROC- which test is better?

Mortality for men age 60-69

ROC Curve

AUC

GLYCOHEMOGLOBIN  &56
FRUCTOSAMINE  &37
GLUCOSE  &18
Reference Line

ROC- value of unrelated tests?

Women age 20-59

ROC Curve

AGE  711
GGT  649
Cholesterol-HDL index  584
Reference Line
ROC Test values

Assessing an article

- Did the (Cox) multivariate analysis:
  - Compare meaningful groups or just extreme values designed to obtain a positive publishable result
  - Have CIs that support the conclusion?
  - Include all important variables or at least discuss the issue
  - Include collinear variables without discussion
- Did the ROC curve ignore an important variable?
- Are the results biologically plausible?