Survival modeling
Survival modeling

• Survival data analysis
• Type of censoring
• Kaplan-meier curves
  – Comparison of curves
  – Log-rank test, Wald test
• Application in R
Survival data analysis

- Statistical modeling focused on time until an event occurs
- Possible events
  - Time to death
  - Time to onset of a disease
  - Time to relapse
  - Time to recurrence
  - Length of stay in a hospital
Survival data

• Survival data present in
  – Clinical trials
  – Cohort studies
  – Retrospective analysis

• Typically survival data are not fully observed but are *censored*
A gene-expression signature to predict survival in breast cancer across independent data sets

A Naderi¹,⁷, AE Teschendorff¹,⁷, NL Barbosa-Morais¹,⁶, SE Pinder², AR Green³, DG Powe³, JFR Robertson⁴, S Aparicio¹, IO Ellis⁵, JD Brenton¹ and C Caldas¹

¹Cancer Genomics Program, Department of Oncology, Hutchison/MRC Research Center, University of Cambridge, Cambridge, UK; ²Department of Pathology, Hutchison/MRC Research Center, University of Cambridge, Cambridge, UK; ³Department of Molecular Medical Sciences, Nottingham City Hospital NHS Trust and University of Nottingham, Nottingham, UK; ⁴Department of Medical and Surgical Sciences, Nottingham City Hospital NHS Trust and University of Nottingham, Nottingham, UK; ⁵Department of Histopathology, Nottingham City Hospital NHS Trust and University of Nottingham, Nottingham, UK and ⁶Faculty of Medicine, Institute of Molecular Medicine, University of Lisbon, Lisbon, Portugal

Introduction

Microarray expression profiling has shown promise for prognostication of breast cancer. van’t Veer et al. (2002) identified a 70-gene-expression signature, which predicted the outcome of pre-menopausal patients with more accuracy than conventional prognostic indicators and validated their signature in a follow-up study (van de Vijver et al., 2002). However, this validation was imperfect as the training and validation cohorts had overlapping patients and external validation using...
Survival curves

![Survival curves graph]

- **Overall Survival**
- **Time since diagnosis (Years)**
- **Group 1**
- **Group 2**
- **Group 3**
The survival time

• Continuous valued
• Incomplete
  – For some subjects we will know their survival is at least $t$
  – For other subjects we know the exact survival time
• Incomplete responses are called censored
• $\geq 0$
Censoring

- Right vs. left censoring
Censoring

- Right vs. left censoring
Right censoring

- Right censoring or suspended data
- Units that did not fail
- A unit is still running but lost to follow-up
- The event is to the right of our timing
Right censoring

- Right censoring or suspended data
- Units that did not fail
- A unit is still running but lost to follow-up
- The event is to the right of our timing

[Diagram showing units with right censoring and an event that did not happen but is too far to the right]
Interval censoring

- Uncertainty on the exact time a unit failed
- Mostly when units are not constantly monitored
- Unit failed between x and y time
Left censoring

- Similar to interval censoring
- Failure time is known to be before a certain time.
- Identical to interval censoring

![Diagram showing data with left censoring](image)
Right censoring

- In biomedicine mostly right censoring
- Censoring mechanism must be independent of the survival mechanism
- Multiple reasons
  - Study ends
  - Lost to follow-up
  - Withdrawal from study
Right censoring
Right censoring

Event did not happen but is to the right
Right censoring

Subjects dropped out before event happened
Example survival data

- Go to http://gdac.broadinstitute.org/
- Pick a cancer
  - E.g. ESCA – Esophageal carcinoma
- Click Browse in Data tab
- Download “Merge_clinical” under clinical
- Open <CancerCode>.clin.merged.txt
  - Tab delimited file
  - Patients in columns
  - Variables in rows
Example survival data

- Open `<CancerCode>.clin.merged.txt`
  - `patient.bcr_patient_barcode`: TCGA identifier of patients
Example survival data

• Open `<CancerCode>.clin.merged.txt`
  – `patient.bcr_patient_barcode`: TCGA identifier of patients

TSS=center patient was collected
Example survival data

- Open `<CancerCode>.clin.merged.txt`
  - `patient.bcr_patient_barcode`: TCGA identifier of patients

Summarize duplicate patients

Only select patients with sample code “01”
Example survival data

- Open `<CancerCode>.clin.merged.txt`
  - Survival data encoded in three variables
    - `patient.days_to_death`
    - `patient.days_to_last_followup`
    - `patient.vital_status`
## Example survival data

<table>
<thead>
<tr>
<th>Patient</th>
<th>Patient days to death</th>
<th>Patient days to last followup</th>
<th>Patient vital status</th>
</tr>
</thead>
<tbody>
<tr>
<td>bcr_patient_barcode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gf</td>
<td>784</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gg</td>
<td>610</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gh</td>
<td>951</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gi</td>
<td>435</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gj</td>
<td>1781</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9go</td>
<td>494</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gq</td>
<td>128</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-2h-a9gr</td>
<td>987</td>
<td>NA</td>
<td>dead</td>
</tr>
<tr>
<td>tcga-ic-a6re</td>
<td>NA</td>
<td>234</td>
<td>alive</td>
</tr>
<tr>
<td>tcga-ig-a4qs</td>
<td>NA</td>
<td>8</td>
<td>alive</td>
</tr>
<tr>
<td>tcga-ig-a7dp</td>
<td>NA</td>
<td>34</td>
<td>alive</td>
</tr>
<tr>
<td>tcga-jy-a6f8</td>
<td>NA</td>
<td>3714</td>
<td>alive</td>
</tr>
<tr>
<td>tcga-jy-a6fb</td>
<td>NA</td>
<td>1837</td>
<td>alive</td>
</tr>
</tbody>
</table>
## Example survival data

<table>
<thead>
<tr>
<th>Patient</th>
<th>bcr_patient_barcode</th>
<th>Time to event or censoring</th>
<th>Event=Dead</th>
<th>Censoring=Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>tcga-2h-a9gf</td>
<td>tcga-2h-a9gf</td>
<td>784</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gg</td>
<td>tcga-2h-a9gg</td>
<td>610</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gh</td>
<td>tcga-2h-a9gh</td>
<td>951</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gi</td>
<td>tcga-2h-a9gi</td>
<td>435</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gj</td>
<td>tcga-2h-a9gj</td>
<td>1781</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9go</td>
<td>tcga-2h-a9go</td>
<td>494</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gq</td>
<td>tcga-2h-a9gq</td>
<td>128</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-2h-a9gr</td>
<td>tcga-2h-a9gr</td>
<td>987</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>tcga-ic-a6re</td>
<td>tcga-ic-a6re</td>
<td>234</td>
<td>alive</td>
<td></td>
</tr>
<tr>
<td>tcga-ig-a4qs</td>
<td>tcga-ig-a4qs</td>
<td>8</td>
<td>alive</td>
<td></td>
</tr>
<tr>
<td>tcga-ig-a7dp</td>
<td>tcga-ig-a7dp</td>
<td>34</td>
<td>alive</td>
<td></td>
</tr>
<tr>
<td>tcga-jy-a6f8</td>
<td>tcga-jy-a6f8</td>
<td>3714</td>
<td>alive</td>
<td></td>
</tr>
<tr>
<td>tcga-jy-a6fb</td>
<td>tcga-jy-a6fb</td>
<td>1837</td>
<td>alive</td>
<td></td>
</tr>
</tbody>
</table>
Example survival data

- Clinical data is messy
- Quality control
  - Missing values
  - Check if days-to-followup < days-to-death
  - Check if value in at least one of days-to-followup or days-to-death
  - Etc.
Survival time modeling

- $T$ is the time to event
- The survival function is then
  - $S(t) = \Pr(T > t)$
  - Usually $S(0) = 1$ and $S(\text{Infinity}) = 0$
  - $S(t)$ is a decreasing function
Survival curve

- $S(t)$ can be interpreted as a proportion.
- It can be estimated as the observed proportion of subjects surviving time point $t$
  - $S(t) = P(\text{Outcome} > t)$
  - $S(t) = (\# \text{ subjects surviving } t) / N$
Survival curve

- $S(t)$ can be interpreted as a proportion
- It can be estimated as the observed proportion of subjects surviving time point $t$
  - $S(t) = P(\text{Outcome} > t)$
  - $S(t) = \frac{\text{(# subjects surviving } t)}{N}$
Survival curve
Survival curve

- In practice $S(t)$ can be estimated as step function
- Steps only when events are observed
- Step size at each time point =
  - $(\# \text{ subjects with event } t)/N$
Survival curve

Overall Survival vs Time since diagnosis (Years)

Number at risk:

<table>
<thead>
<tr>
<th>Group</th>
<th>29</th>
<th>18</th>
<th>10</th>
<th>4</th>
<th>3</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>32</td>
<td>27</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group 3</td>
<td>31</td>
<td>29</td>
<td>28</td>
<td>24</td>
<td>19</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Survival curves with confidence intervals
Hazard function

• Hazard function is denoted \( h(t) \)
  – Instantaneous potential per unit time for the event to occur given that the individual has survived until time \( t \)

• Hazard function is the opposite of the survival function
Hazard function

- Hazard function represents failure
  - The hazard is a rate rather than a probability
  - Ranges from 0 – infinity.

\[
\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt \mid T \geq t\}}{dt}.
\]
Hazard function

\[ \lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}. \]

- Numerator represents the conditional probability that the event happens in the interval \([t, t+dt)\) given that it has not happened before.
- Denominator is the width of the interval.
- Rate of event occurrence per unit of time.
- Limit gives instantaneous rate of occurrence.
Hazard function

• With some math one can show that:

\[
\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt \mid T \geq t\}}{dt}.
\]

\[
\lambda(t) = \frac{f(t)}{S(t)}.
\]

• Showing again that hazard is the inverse of survival
Summary

• Hazard function
  – Risk of event happening in a time interval after time t, given that patient has survival to time t
  – h(t)

• Survival function
  – Probability that patient will have survival time >=t
  – S(t)
Regression for survival data

• Most frequent used model is Cox model
• \( h(t,X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \)
• With \( X = (X_1, X_2, \ldots, X_p) \) predictor variables e.g. gene expression
Regression for survival data

• \( h(t,X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \)

• Product of two quantities
  – \( h_0(t) \) baseline hazard
  – Exponential of the sum \( \beta_i \) and \( X_i \)

• Baseline hazard is not specified
  – Semi-parametric model
  – Reason for cox model being popular
Regression for survival data

- Cox models effect of predictors on hazard rate but leaves baseline hazard unspecified
- Does not assume knowledge of absolute risk rather estimates relative risk
  - Proportional hazards modeling
- Maximum likelihood to estimate $\beta_i$ exists
Covariates

• Often assumed independent of time if they are unlikely to change
• Not changing over time
  – Smoking
  – Age
  – Weight
  – Etc.
• Values set at t=0
Survival curves

• Kaplan Meier survival curves
  – Visualization of survival analysis
  – Complimentary to Cox proportional hazards modeling
  – Non-parametric modeling
  – KM estimates a probability at each event time $t_i$

\[ p_i = \frac{\text{number of events at time } t_i}{\text{number of subjects at time } t_i} \]
Kaplan Meier curve comparison

• Compare whether two groups are significantly different from each other
• Compares two survival curves for statistical difference
• Log-rank or Mantel-Haenszel test
• Needs grouped data, no continuous valued variables
Survival curve

Overall Survival

Time since diagnosis (Years)

Number at risk

<table>
<thead>
<tr>
<th>Group</th>
<th>29</th>
<th>18</th>
<th>10</th>
<th>4</th>
<th>3</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>32</td>
<td>27</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>31</td>
<td>29</td>
<td>28</td>
<td>24</td>
<td>19</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

n=92

$P_{1-3} < 0.001^*$
Survival curves with confidence intervals

Cox-ranked signature (70 genes)

Overall Survival

p = 2.7e-07
HR = 5.97 (3.0-11.9)

Years

n=13/85
n=22/37

Good
Poor
Kaplan Meier comparison

• For continuous values, need to make groups
  – E.g. gene expression or gene signature score
• Split across median, tertiles, quantiles etc.
• Possible to optimize split
  – Find best possible threshold with largest survival difference
  – Danger of overfitting
  – Need train and test set
Problem set

- R is your friend for survival analysis
- Packages
  - survival
  - maxstat
  - Glmnet

Few other programming environments have the same extensive implementations for statistical models for survival analysis
Next lecture

• Multiple testing
• Classifiers
  – Bayesian modeling
  – Kernel methods
  – Regularized regression
  – Sparse regression