# Survival analysis - Overview

1. What is survival analysis?

* Also known as **time-to-event analysis**
* A branch of statistics that studies the amount of time it takes before a particular event of interest occurs
* Applications: Insurance companies – predict the death of the insured and estimate other important factors (e.g., policy cancellations, non-renewals), and how long it takes to file a claim -> calculate insurance premiums, lifetime value of customers
* Initially developed in biomedical sciences to understand the onset of certain diseases – now used in engineering, insurance, etc.

1. Understanding survival analysis

* Mainly comes from the medical and biological disciplines – study rates of death, organ failure, onset of diseases
* Can also apply to positive events (e.g., how long it might take someone to win the lottery if they play it every week)
* Adapted to biotechnology sector, economics, marketing, machine maintenance, engineering, etc.
* Uses the **hazard rate** to measure the odds / chances of an item or system failing – dependent upon the amount of time the item / system has been in use
* Survival analysis in insurance:
* Use to outline the incidence of death at different ages given certain health conditions
* Compute the prob of whether policyholders will outlive their life insurance coverage
* Providers can calculate an appropriate insurance premium (the amount each client is charged for protection) by also taking into account thee value of the potential customer payouts under the policy
* E.g.: estimate how long it will take drivers from a particular zip code to have an auto accident, based on location, age, type of insurance, time since last claim filed

1. Advantages & disadvantages

* Other more common statistical methods: regression analysis – commonly used to determine how specific factors influence the price movement of an asset
* Problem: linear regression often makes use of both positive and negative numbers
* **Survival analysis deals with time, which is strictly positive**
* Linear regression is not able to account for censoring – survival data that is not complete for various reasons (**right-censoring** – the subject has not yet experienced the expected event during the studied time period)
* Main benefit of survival analysis: better tackle the issue of censoring – main variable, other than time, addresses **whether the expected event happened** or not

# Survival Analysis Part I: Basic concepts and first analyses

1. Introduction

* Cancer studies: Main outcome under assessment – time to an event of interest (survival time)
* It is usual that at the end of follow-up, some of the individuals have not had the event of interest -> true time to event is unknown
* Survival data – rarely normally distributed – skewed & comprise typically of many early events and relatively few late ones
* Survival analysis necessary
* Most survival analyses in cancer use some or all: Kaplan-Meier plots, logrank tests, Cox (proportional hazards) regression

1. Types of ‘event’ in cancer studies

* Medical studies**: time to death** is the event of interest
* However, in cancer, another important measure: **time between response to treatment** and recurrence or **relapse-free survival time** (disease-free survival time)
* What the event is & when the period of observation start & finishes
* E.g.,: relapse in the time period between a confirmed response and the first relapse of cancer

1. Censoring

* Specific difficulties of survival analysis: only some individuals have experienced the event and survival times will be unknown for a subset of the study group
* Censoring, may arise in the following ways:
* A patient has not (yet) experienced the relevant outcome (e.g., relapse or death) by the time of the close of the study
* A patient is lost to follow-up during the study period
* A patient experiences a different event that makes further follow-up impossible
* **Right-censoring**: The event (assuming it occurs) **is beyond the end of the follow-up** period
* **Left-censoring** can also occur if we observe the presence of a state or condition but **do not know where it began**
* **Interval censor**: individuals come in and out of observation
* The feature of censoring means that special methods of analysis are needed

1. Survival and hazard

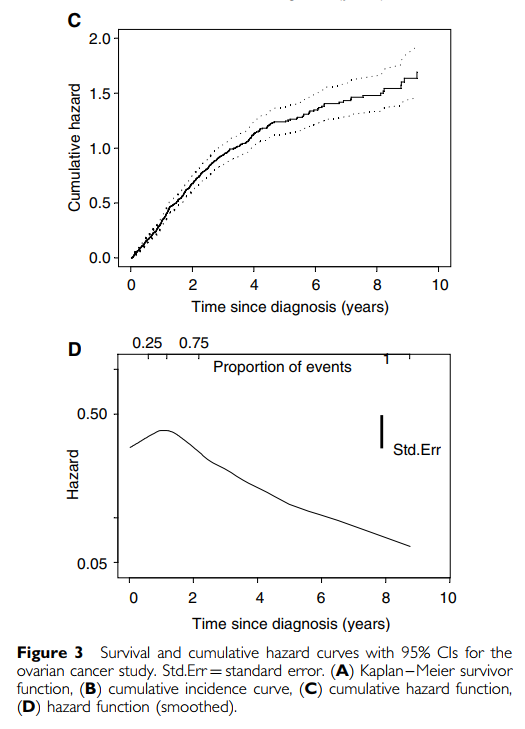
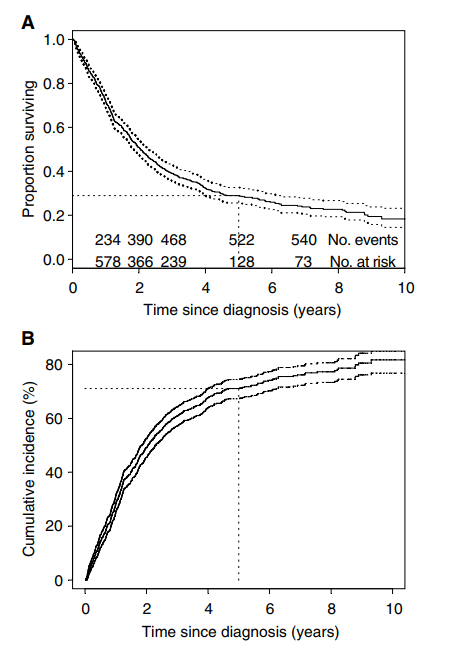
* Survival data generally described and modelled in terms of **2 related probabilities: survival and hazard**
* **Survival probability (survival function) - S(t)**: the probability that an individual survives from the time origin (e.g., diagnosis of cancer) to a specified future time t.
* Survival prob for different values of t provide crucial summary info from time to event data
* Describe directly survival experience of a study cohort
* Focuses on **not having an event**
* **Hazard – h(t) or λ(t):** the probability that an individual who is under observation at a time t has an event at that time
* Instantaneous event rate for an individual who has already survived to time t
* Focuses on the **event occurring**
* Insight into the conditional failure rates -> help specifying a survival model
* Hazard relates to the incident (current) event rate, while survival reflect the cumulative non-occurrence

1. Kaplan-Meier survival estimate

* Survival prob can be estimated nonparametrically from observed survival times, both censored and uncensored, using the KM (product-limit) method (Kaplan and Meier, 1958)
* Suppose k patients have events in the period of follow-up at distinct times
* Events are assumed to **occur independently** of one another
* The prob of **surviving from one interval to the next** may be **multiplied** together to give the **cumulative survival prob**
* **The prob of being alive at time**  is calculated by:

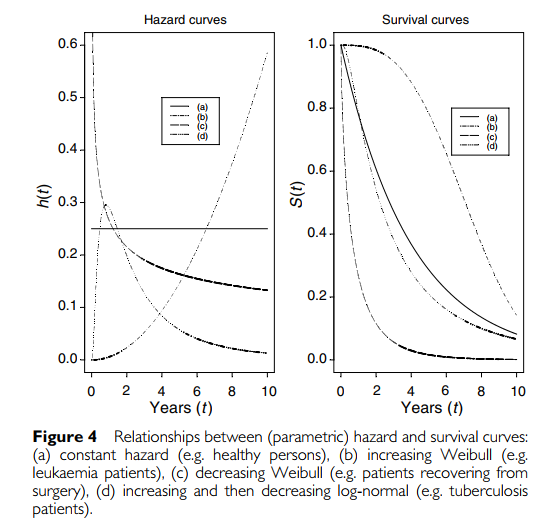
Where:

* : prob of being alive at time
* : number of patients alive just before
* : number of events at
* S(0) = 1
* **is constant between times of events**, and therefore the estimated prob is a step function that changes value only at the time of each event
* This estimator allows each patient to contribute info to the calculations for as long as they are known to be event-free
* In case of no censoring (i.e., every individual is to experience the event), this estimator would simply reduce to the ratio of the number of individuals event-free at time t divided by the number of people entering the study
* **Cumulative incidence at a time point** = 1 – survival prob.
* Confidence intervals for the survival prob:
* The KM survival curve, a plot of KM survival prob against time provides a useful summary of the data that can be used to estimate measures such as median survival time (large skew in distribution of most survival data -> mean is not often used)
* In practice, there are usually patients who are lost to follow-up or alive at the end of follow-up, and confidence limits are often wide at the tail of the curve, making meaningful interpretations difficult
* E.g.:



1. Hazard and cumulative hazard

* Relationship between
* Unimportant for routine survival analyses as it is incorporated into most statistical computer packages
* If either S(t) or h(t) is known, the other is automatically determined >< Unlike S(t), there is no simple way to estimate h(t)
* Instead, a quantity called the **cumulative hazard**  is commonly used, defined as the **integral of the hazard**, or the area under the hazard function between times 0 and t, and differs from the log-survivor curve only be sign, i.e.,
* **Interpretation of :** cumulative force of mortality, or the number of events that would be expected for each individual by time t if the event were a repeatable process
* is used as an intermediary measure for estimating and as a diagnostic tool in assessing model validity
* Nonparametric method for estimating : Nelson-Aalen estimator, from which it is possible to derive an estimate of by applying a kernel smoother to the increments. Cox (1979) – another method to estimate the hazard based on order statistics but similar in spirit to the previous method
* Another approach for estimating hazard: assume that the survival times follow a specific mathematical distribution
* Constant hazard rate over time (analogous to exponential distribution of survival times)
* Strictly increasing / decreasing hazard rates based on a Weibull model
* Combination of decreasing and increasing hazard rates using a log-normal model

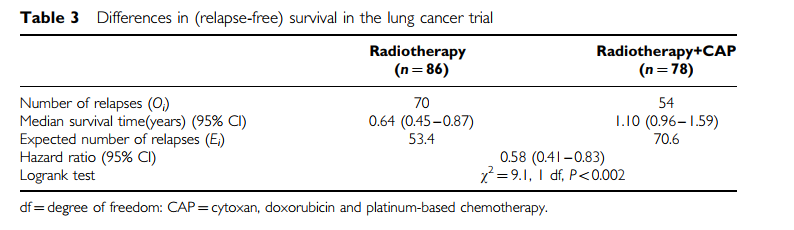


1. Nonparametric tests comparing survival

* **Compare survival in 2 or more groups** of patients - logrank test (Peto et al, 1977)
* The method calculates **at each event time, for each group, the number of events one would expect** since the previous event if there were **no difference between the groups.** Values are then summed over all event times to give the **total expected number of events in each group**, say
* The logrank test **compares the observed number of events**, say for treatment group i, **to the expected number** by calculating the test statistic
* This value is **compared to a distribution** with (g-1) degrees of freedom, where g is the number of groups
* **P-value** may be computed to calculate the statistical significance of the differences between the complete survival curves
* If the groups are **naturally ordered** -> consider the possibility that there is a **trend in survival** across them, for example, age groups or stages of cancer
* Calculating and for each group on the basis that survival may increase or decrease across the groups -> more powerful test
* For the new and , the test statistic for trend is compared with distribution with one degree of freedom
* **Only 2 groups**: logrank test is testing the **null hypothesis that the ratio of the hazard rates in the 2 groups = 1**
* **Hazard ratio (HR)** – measure of the relative survival experience in the 2 groups and may be estimated by

Where is the estimated relative (excess) hazard in each group i.

* Confidence interval for HR
* HR has similar interpretation of the strength of effect as a risk ratio
* HR = 1 -> no difference in survival
* In practice, better to estimate HRs using a regression modeling technique, using Cox regression



* Other nonparametric tests include methods to compare median survival times, but comparing confidence intervals for each group is not recommended (Altman and Bland, 2003).

1. Some key requirements for the analysis of survival data

* **Uninformative censoring**
* Standard methods to analyze survival data with censored observations are **valid only if the censoring is ‘noninformative’** -> censoring carries no prognostic info about subsequent survival experience. Those who are censored because of loss to follow-up at a given point in time should be as likely to have a subsequent event as those individuals who remain in the study.
* However, when the number of patients lost to follow-up is small, very little bias is likely to result from applying methods based on noninformative censoring
* **Length of follow-up**
* Time to event studies must have **sufficient follow-up to capture enough events** -> sufficient power to perform appropriate statistical test
* Proposed length of follow-up – based primarily on the severity of the disease or prognosis of participants
* E.g.,: lung cancer trial – 5-year follow up; breast cancer patients – need longer follow-up duration
* An indicator of length of follow-up – median follow-up time, calculated from follow-up among the individuals with censored data
* Reverse KM estimator – more robust measure: Event indicator reversed so that the outcome of interest becomes being censored
* **Completeness of follow-up**
* Unequal follow-up between different groups (e.g., treatment arms) may bias the analysis
* **Disparities in follow-up** caused by differential drop-out between arms of a trial or different subgroups in a cohort study need to be investigated
* **Cohort effect on survival**
* Assumption of **homogeneity of treatment and other factors during the follow-up period** >< in a long-term observational study of patients of cancer, the case mix may change over the period of recruitment, or there may be an innovation in ancillary treatment.
* KM method assumes that the **survival prob are the same for subjects recruited early and late** in the study
* Between-center differences
* Consistency between the study methods in each center

1. Need for survival analysis adjusting for covariates

* When comparing treatments in terms of survival, it is often sensible to adjust for patient-related factors (covariates or confounders), which could potentially affect the survival time of a patient
* E.g., suppose that despite the treatment being randomized in the lung cancer trial, older patients were assigned more often to the radiotherapy alone group -> this group will have worse baseline prognosis -> underestimated efficacy compared to the combination treatment
* Confounding between treatment and age
* Sometimes, we want to determine the prognostic ability of various factors on overall survival
* Multiple prognostic factors can be adjusted for using multivariate modeling

1. Summary

* Survival analysis: a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs
* Censoring -> a proportion of survival times of interest will often be unknown
* Assumption: patients who are censored have the same survival prospects as those who continue to be followed -> uninformative censoring
* Survival data – generally described and modeled in terms of 2 related functions – survival function & hazard function
* Survivor function – prob that an individual survives from the time of origin to some time beyond time t
  + directly describes the survival exp of a study cohort
  + usually estimated by the KM method
  + test for differences between survival curves for groups – logrank test
* Hazard function – instantaneous potential of having an event at a time, given survival up to that time
  + Used primarily as a diagnostic tool or for specifying a mathematical model for survival analysis
* In comparing treatments or prognostic groups in terms of survival, necessary to adjust for patient-related factors that could potentially affect the survival time of a patient
* Multivariate survival analysis (a form of multiple regression) provides a way of doing this adjustment

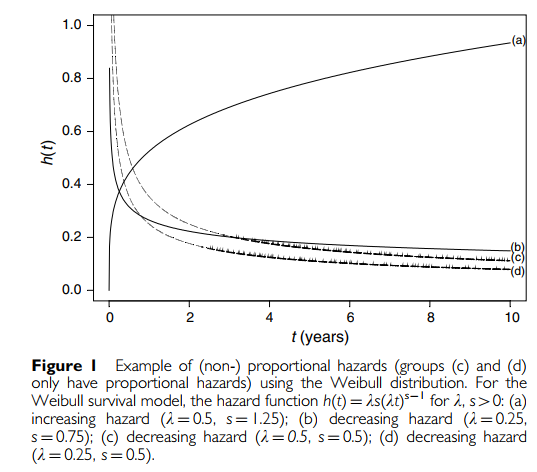
# Survival analysis part 2: Multivariate data analysis – Intro to concepts and methods

1. The need for multivariate statistical modeling

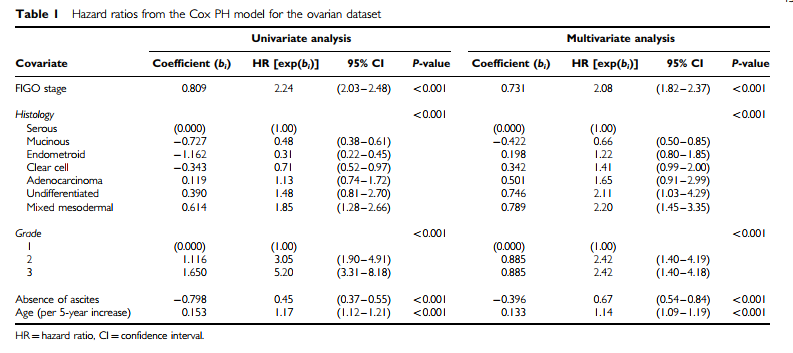
* Previous paper: construction of survival curves for different patient groups (KM), and logrank test to investigate differences between them
* Univariate analysis – survival wrt factor under investigation, ignoring impact of others
* In clinical investigations, common to have a situation where several (known) quantities or covariates, affect the patient prognosis (e.g., genotype, age)
* Desirable to adjust for the impact of others
* Logrank test – provides p-value for the differences between the groups >< no estimate of the actual effect size (not a clinical assessment of the factor’s impact)
* Statistical model – improves on these methods by allowing survival to be assessed **with respect to several factors simultaneously** and offers **estimates of the strength of effect for each constituent** factor
* **Assumptions**:
* Survival times are independent of each other
* Censoring occurs solely as right-censoring
* Censoring are uninformative
* Focus on covariates measured at the time of entry to the study:
* Continuous (e.g. age, tumor size)
* Binary (e.g., gender)
* Unordered categorical (e.g., histology)
* Ordered categorical / ordinal (e.g., performance status, FIGO stage)
* Methods – 2 broad categories:
* Proportional hazard approaches (incl. semi-parametric Cox model and fully parametric approaches)
* Accelerated failure time models

1. **Cox (semi-parametric) proportional hazards model**

* Most commonly used multivariate approach for analyzing survival time data in medical research
* Survival analysis regression model
* Describes the relation between the event incidence, as **expressed by the hazard function and a set of covariates**
* The hazard function is dependent on a set of p covariates , whose impact is measured by the size of the respective coefficients
* The term - baseline hazard: value of the hazard if all the = 0
* **Hazard may vary over time**
* The baseline hazard function is estimated nonparametrically -> survival times are **not assumed to follow a particular statistical distribution**
* Essentially a **multiple linear regression of the logarithm of the hazard** on the variables , with the baseline hazard being an ‘intercept’ term that varies with time
* Covariates act multiplicatively on the hazard at any point in time
* **Key assumption of the PH model**: the hazard of the event in any group is a constant multiple of the hazard in any other
* The hazard curves for the groups should be proportional and cannot cross

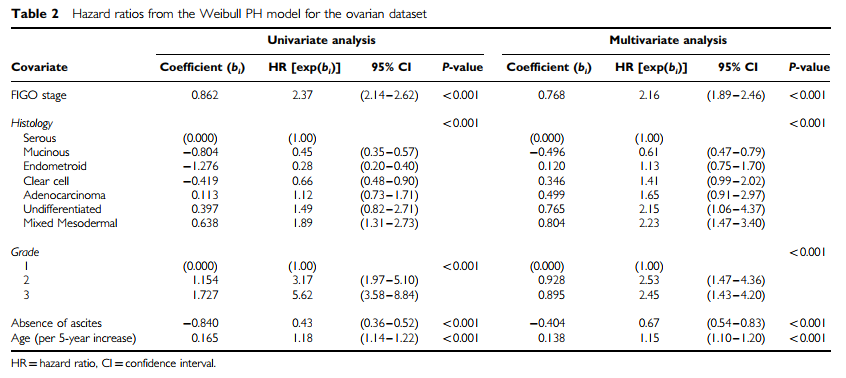


* **Proportionality** implies that the quantities are called hazard ratios
* **Hazard ratio > 1** -> covariate positively associated with the event prob & negatively associated with the length of survival
* Proportionality assumption is often appropriate for survival time data but it is important to verify that it holds
* Example: Cox PH model fitted to ovarian cancer data



1. **Parametric PH model**

* A class of models **similar in concept and interpretation** to the Cox PH model
* Key difference: the **hazard is assumed to follow a specific statistical distribution** when a fully parametric PH model is fitted to the data
* Hazard ratios have the same interpretation, whether derived from a Cox or a fully parametric regression model
* Proportionality of hazards is still assumed
* Choosing **different hazard functions -> different parametric PH models**
* The choice of hazard distribution determines that of the survival
* Models commonly applied (e.g., **Exponential, Weibull, Gompertz**) take their names from the distribution that the survival times are assumed to follow
* Most distinguishing features between them are in the hazard function
* E.g.: Parametric models fitted to the ovarian cancer data



1. **Comparison of the two PH approaches**

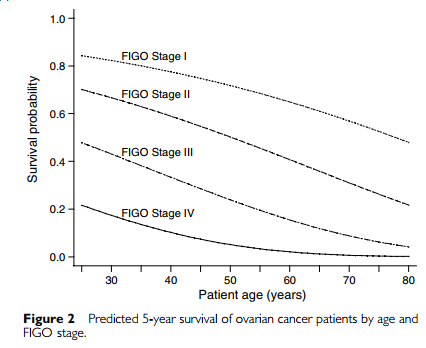
* **Parametric models** – main drawback: need to **specify the distribution** that most appropriately mirrors actual **survivor times**
* May be difficult to identify
* In case a suitable distribution is found, parametric model is more informative than Cox model
* Straightforward to derive the hazard function and obtain predicted survival times
* Slightly more efficient – more precise estimates (smaller standard errors)
* Results from Cox or parametric PH models may be compared directly
* For either method to be valid:
* **Covariate effect needs to be at least approx. constant** throughout the duration of the study
* **Proportionality assumption** must hold

1. **Interpreting the PH model: Beyond the hazard ratio**

* In addition to the ratio of two hazards, predicted survival proportion at any given point in time for a particular risk group can be derived.
* **Survival proportion for a given risk group at any time,** :

is the baseline survival (when all covariates = 0) and

* Baseline survival at a given time -> predicted survival prob for patients with any specified covariate values
* E.g.:

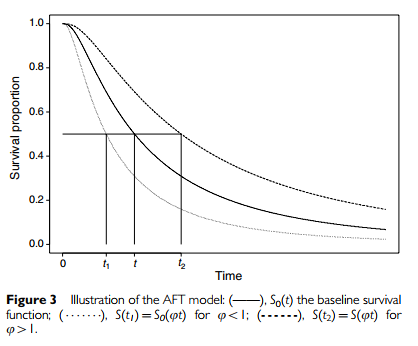


1. **Accelerated failure time models**

* A different type of model used for analysis of survival time data

Where is the baseline survivor function and  **is an ‘acceleration factor’** that depends on the covariates

* Principle: the effect of a covariate is to **stretch or shrink the survival curve along the time axis** by a constant relative amount



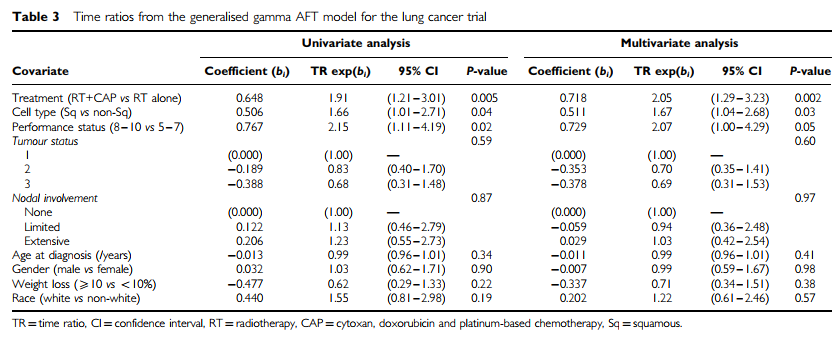
* **AFT model is commonly rewritten as being log-linear with respect to time**

Where is a measure of (residual) variability in the survival times.

* Survival times can be seen to be multiplied by a constant effect under this model specification, and the **exponentiated coef, , are referred to as time ratios**.
* Time ratio > 1 – ‘slows down’ or prolongs time to the event
* Time ratio < 1 – accelerates time to the event
* **Survival time follows a Weibull distribution** -> AFT and PH models are the same

>< AFT models differs from PH models in terms of their interpretation of effect sizes as time ratios as opposed to hazard ratios

* Other quantities (e.g., projected survival prob) may be derived
* **Assumptions**
* Appropriate choice of **statistical distribution of survival times** (e.g., log-normal, log-logistic, generalized gamma, Weibull) needs to be made
* **Covariates effects** are constant and multiplicative on the timescale (i.e., covariate impacts **on survival by a constant factor**)
* E.g.: Parametric AFT models fitted to the lung cancer trial data



1. Which model should we use: PH vs. AFT?

* From statistical viewpoint, an obvious way to choose between the 2 model types is to fit a type that is in keeping with the data
* However, when either type of model may appear to fit the data adequately, the choice of model may be influenced by other factors
* E.g.: if other studies of a similar nature had all used Cox regression and reported results as hazard ratios, one may be tempted to follow suit to aid comparability
* Parametric approach offers more in the way of predictions
* AFT allows the derivation of a time ratio -> more interpretable than a ratio of 2 hazards >< unfamiliar in medical research

1. Other approaches

* **Stratified survival analysis**
* More straightforward way to incorporate covariates into a survival analysis
* E.g.: Suppose the covariate of primary interest is treatment but we wish to control for the clinical stage of the tumor when comparing
* Survival in each treatment group can be compared within each stage of disease (strata) by the logrank or some other methods
* Differences within each stratum are then combined to given an overall comparison of treatments that has been adjusted for the stage
* Strength: Simplicity – logrank test is nonparametric -> few distributional assumptions made + interpretation straightforward
* Limitation: only applicable when the covariate is categorical (or with continuous variables that have been arbitrarily categorized). Does not perform well with several covariates – number of individuals in each stratum quickly becomes too small for reasonable comparisons
* Does not quantify the strength of effect of each variable or even offer a p-value for factors other than the one of primary interest
* Not generally regarded as a formal statistical model
* Can be use where a very small number of covariates are to be considered, if only as an exploratory method of analysis
* **Aalen’s additive model**
* Assume that covariates impact additively upon a (unknown) baseline hazard, but the effects are not constrained to be constant, allowed to vary freely over time

where is the hazard, is the baseline hazard and the are coef (may change in magnitude and sign over time)

* Not straightforward to estimate nonparametrically -> Cumulative baseline hazard is used and the regression coef that are actually estimated from the data are also the cumulative (additional) hazard
* Representing the effect – graph against time
* The further is from 0 at time t, the greater the effect the covariate has had on the hazard over the course of the study up to t
* Values of – absolute increase in hazard at time t – are not actually observed, >< relative size may be inferred from the slope of the line
* Informal assessment of the adequacy of the proportional hazards assumption in the Cox model
* Lack easy interpretation - coef are not easy to understand, changing repeatedly overtime, offering no single quantifiable effect size
* Aalen plots are essentially the only manner with which to interpret the effect sizes
* Classification trees and artificial neural networks

1. Discussion

* Statistical models – ability to assess several covariates simultaneously
* Strengths of the stratified logrank rest and other such methods – simplicity & fewer parametric assumptions of the data
* Cox model – greater flexibility than parametric alternatives and does not require the direct estimation of the baseline hazard function (i.e., avoids the need to specify the distribution of survival times)

>< Proportionality assumption of hazard needs to be fulfilled for the results to be meaningful

* Other parametric models will produce more precise estimates where the distribution is specified correctly
* Further concern – choice of covariates to include (in next paper)