How do you carry out survival analysis?

Nathan Constantine-Cooke (Vallejos group)

MRC Human Genetics Unit

THE UNIVERSITY of EDINBURGH

@IBDNathan

How do you carry out ~~survival~~ time-to-event analysis?

Nathan Constantine-Cooke (Vallejos group)

MRC Human Genetics Unit

THE UNIVERSITY of EDINBURGH

@IBDNathan

What will we cover?

- What is survival analysis?
- The core fundamentals of survival analysis
- Kaplan-Meier
- The Cox proportional hazards model
- Accelerated failure time models
- Competing risks survival analysis
- Joint modelling of time-to-event and longitudinal outcomes

github.com/nathansam/igmm-survival

What *is* survival analysis?

github.com/nathansam/igmm-survival

What *is* survival analysis not?

Survival analysis is *not* statistical epidemiology

github.com/nathansam/igmm-survival

What *is* survival analysis?

"The analysis of the expected duration of time until one or more events happen"

-Wikipedia et al.

github.com/nathansam/igmm-survival

What *is* survival analysis?

Usually, we are interested in modelling the time to an event of interest, using observable characteristics.

github.com/nathansam/igmm-survival

What *is* survival analysis?

Applications in

- Drug trials
- Precision/personalised medicine
- NHS resource management
- Engineering
- Criminology

github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

Events of interest. What could an event be?

- Death
- Hospital discharge
- Onset of disease
- Clinical improvement or clinical deterioration
- The failure of a mechanical system
- An ex-prisoner re-offending (recidivism)

The core fundamentals of survival analysis

Assumptions for an event of interest

- The event will happen *eventually* for all individuals in a cohort.
- The event is binary. It can't "half happen".
- The event only happens once for an individual.

The core fundamentals of survival analysis

Defining an initial t_0 value is an important consideration.

- t_0 should represent the onset of risk.
- At t_0, all individuals should be at the same risk.
- No one should have experienced the event at, or before, t_0.
- Possible time points for t_0:
    - Birth
    - Hospital admission
    - Diagnosis
    - Medicine first being prescribed

github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

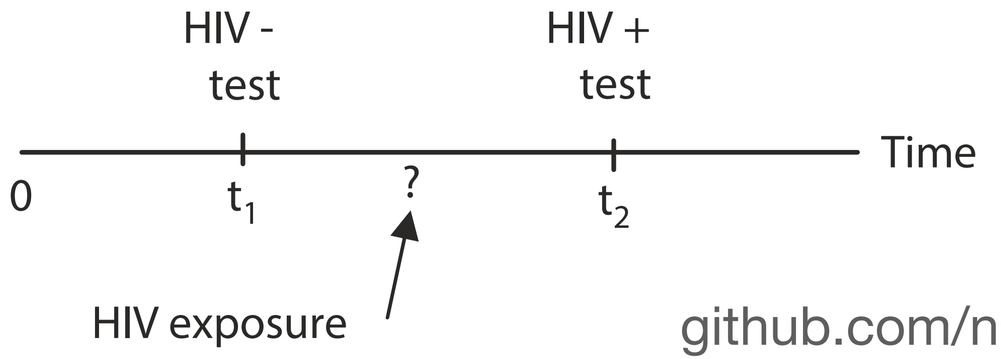
Right-censoring

- In most real-world situations, censoring is encountered in survival analysis.
- This is usually due to:
    - A study ending or reaching maximum follow-up
    - Patient is lost to follow-up
    - Patient withdraws from the study
- These suggestions are all examples of right-censoring.

The core fundamentals of survival analysis

Interval censoring

- We may also encounter interval censoring: where we are only aware of an interval the event occurred in. This is common when a patient has regular tests.

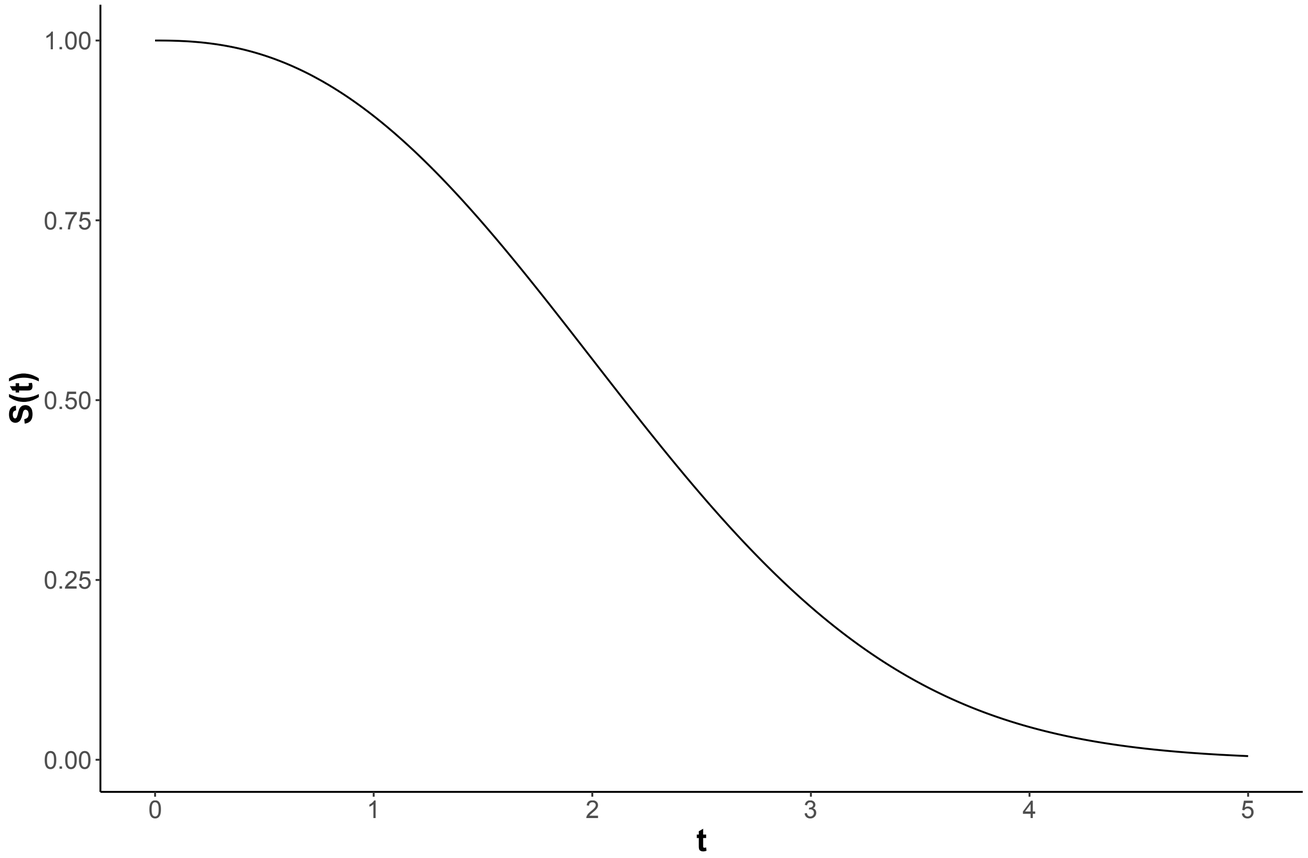


github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

Theoretical Survival functions/ curves

- A survival function gives the probability that a person survives longer than some specified value of time, t.
- Theoretically, this function is a smooth curve.

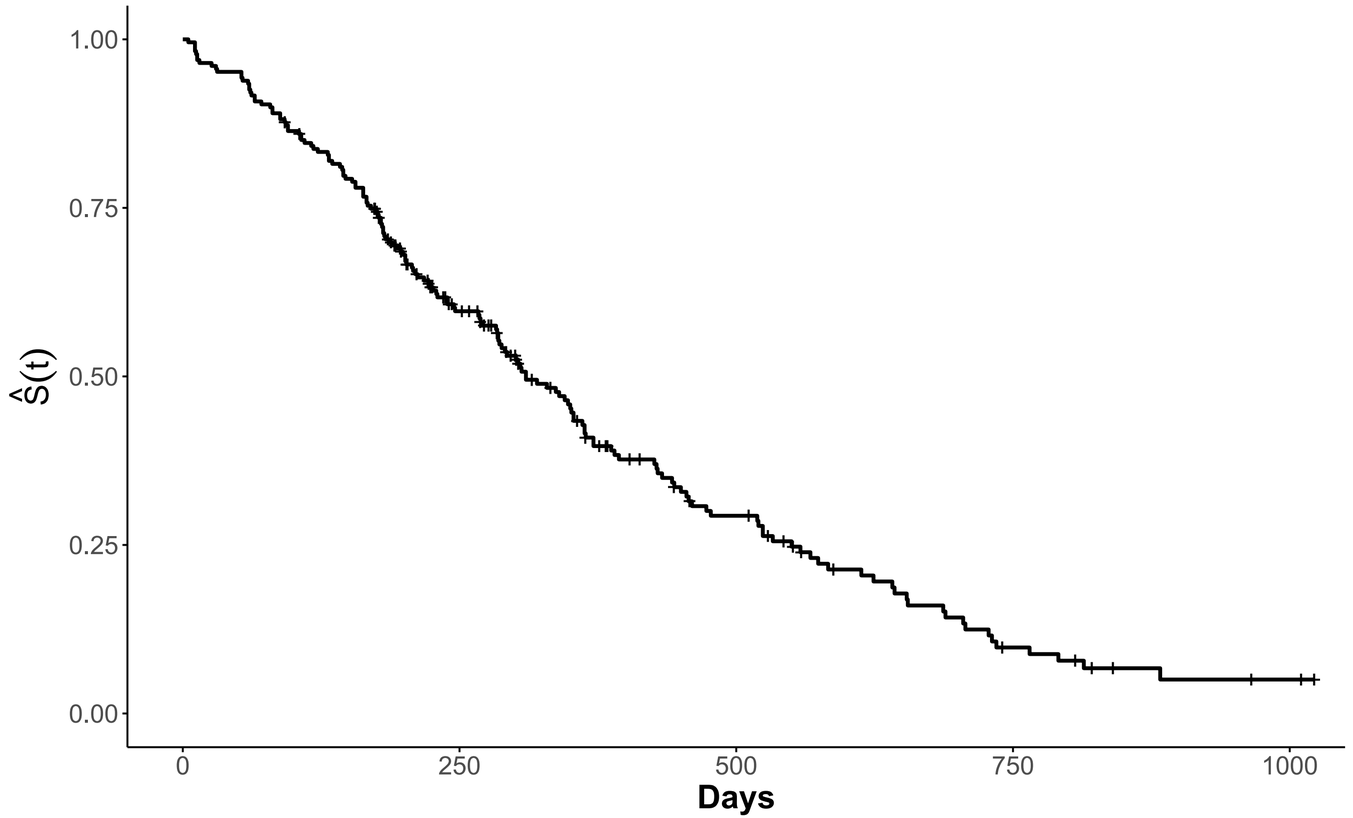


github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

Survival functions/ curves in reality…

- Using actual data usually results in a step function.
- We use vertical bars to indicate censoring.



github.com/nathansam/igmm-survival

The core fundamentals of survival analysis

Hazard functions

- A hazard function, $h(t)$, gives the instantaneous potential for the event to occur at time t, given the subject has survived up to this time.
- A conditional failure rate (rather than a probability).
- We typically focus on the hazard function when mathematically modelling survival data, and also present a model via its hazard function.
- If we have a hazard function, we can derive the survival function and vice versa.

Kaplan-Meier

github.com/nathansam/igmm-survival

Kaplan-Meier

Introduction to Kaplan-Meier

- The Kaplan-Meier statistic is a non-parametric statistic which estimates the true survival function.
- Step-wise.
- Can handle right-censored data.
- Can be used to compare the survival times between groups (such as a placebo group and treatment group.
- Extensively used in survival analysis (58,000 < citations).

github.com/nathansam/igmm-survival

Kaplan-Meier

The Kaplan-Meier estimator

$$S(t) = \prod_{i:t_i \geq t} \left(\frac{n_i - d_i}{n_i} \right)$$

Where d_i is the number of people who experience the event at time t_i

And n_i is the number of people at risk at time t_i (we call this the risk set)

github.com/nathansam/igmm-survival

Kaplan-Meier

Practical example

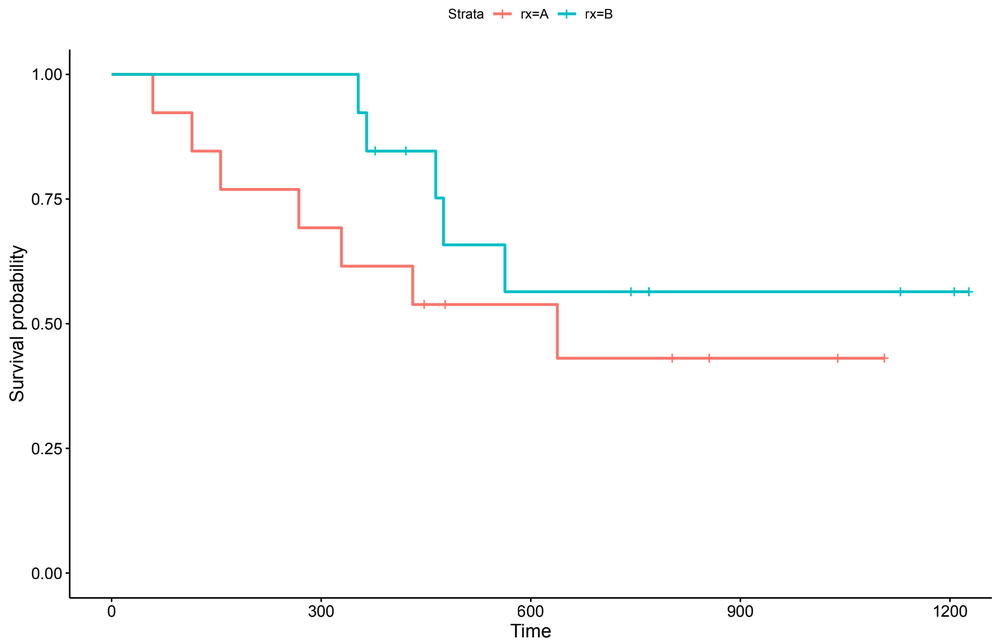
$t_{(f)}$	n	Event	Censored	$S(t_{(f)})$
0	20	0	0	$\frac{20}{20} = 1$
1	20	2	0	$1 \times \frac{18}{20} = 0.9$
2	18	4	2	$0.9 \times \frac{14}{18} = 0.7$
3	12	3	1	$0.7 \times \frac{9}{12} = 0.525$

github.com/nathansam/igmm-survival

Kaplan-Meier

Stratified Kaplan-Meier

- We may wish to investigate if two groups have equivalent survival curves.
- We can draw separate survival curves for each group (ideally with confidence intervals) to visualise how the survival curves vary over time.
- But, we would like to be able to perform a statistical test to empirically determine if there is a difference between the groups.



github.com/nathansam/igmm-survival

Kaplan-Meier

The log-rank test

- We can generate a test statistic and associated p-value via a log-rank test.
- H_0: The survival curves for all groups are the same.
- H_1: The survival curves for all groups are not the same.
- Easy to perform in statistical software (R examples on the Github repo)

github.com/nathansam/igmm-survival

The Cox proportional hazards model

github.com/nathansam/igmm-survival

The Cox proportional hazards model

Introduction to Cox proportional hazards

- A semi-parametric class of models. This means we do not need to choose a specific statistical distribution (such as the Normal distribution)
- Commonly used in medical statistics
- Uses covariates which do not vary with time
- An extension for time-dependent variables exists (but is often inadvisable to use)

The Cox proportional hazards model

The hazard function 1

- The Cox proportional hazards model is typically presented via the hazard function:

$$h(t, X) = h_0(t) \cdot e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}$$

Intercept

Covariate effects

github.com/nathansam/igmm-survival

The Cox proportional hazards model

The hazard function 2

- The baseline hazard is dependent only on time, whilst the second term is dependent only on our covariates and their regression coefficients

$$h(t, X) = h_0(t) \cdot e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}$$

Baseline hazard

Second term

- Interpretation: the baseline hazard is the hazard in the absence of any covariates.

The Cox proportional hazards model

Hazard ratios

- A hazard ratio (HR) is defined as the hazard for one individual/group divided by the hazard for a different individual/group. For example, is a placebo group vs. a treatment group.

$$\text{HR} = \frac{h(t, X = \text{placebo})}{h(t, X = \text{treatment})}$$

- The group with the largest hazard is typically the numerator as hazard ratios ≥ 1 are easier to interpret.

The Cox proportional hazards model

Hazard ratios 2

$$
\begin{aligned}
\text{HR} &= \frac{h_0(t) \cdot e^{(\beta_1 \times 1)}}{h_0(t) \cdot e^{(\beta_1 \times 0)}} \\
&= \frac{e^{\beta_1}}{e^0} \\
&= e^{\beta_1}
\end{aligned}
$$

If x_i is a 0/1 exposure variable, then HR $= e^{\beta_i}$ is the marginal effect size (in terms of HR) of an exposure - provided all other covariates are constant and there are no interaction terms.

The Cox proportional hazards model

Hazard ratio confidence intervals

A 95% confidence interval for the hazard ratio corresponding to the i^{th} covariate can be predicted using

$$\exp\left(\hat{\beta}_i \pm 1.96\sqrt{\hat{Var}\hat{\beta}_i}\right)$$

Where $s_{\hat{\beta}_i} = \sqrt{\hat{Var}\hat{\beta}_i}$

Easy to calculate when there are no interaction effects in the model, much more complicated when there are interaction terms. Fortunately this is also already implemented in most statistical software!

github.com/nathansam/igmm-survival

The Cox proportional hazards model

The proportional hazards assumption.

- HR is assumed to be constant over time. Recall that a hazard ratio does not depend on time.
- If this property is not satisfied then we should consider alternatives such as stratified Kaplan-Meier or accelerated failure time models

The Cox proportional hazards model

Schoenfeld residuals

- We would of course like to be able to statistically test for if the proportional hazards assumption is valid
- A popular method is via Schoenfeld residuals.
- If the proportional hazards assumption is satisfied for a covariate then the Schoenfeld residuals for that covariate will not be related to survival time
- If we rank event times, we can then test the correlation between Schoenfeld residuals and their respective ranks

The Cox proportional hazards model

Parametric proportional hazards

- By specifying a parametric model, for the baseline hazard, we can formulate different kinds of proportional hazards models.
- This is useful for predictive models.
- Popular parametric families used for parametric proportional hazards:
    - Exponential
    - Weibull

github.com/nathansam/igmm-survival

Accelerated failure time models

github.com/nathansam/igmm-survival

Accelerated failure time models

Introduction to AFTs

- Common alternative to Cox proportional hazards
- Fully parametric
- More consistent with theoretical $S(t)$ than Cox (as it is not a step function)
- Hazard & survival function specified
- Results are usually considered more interpretable than Cox. Possible example: patient could be expected to live $15\,\%$ longer if they ceased smoking.

github.com/nathansam/igmm-survival

Accelerated failure time models

The Weibull distribution

Distribution	S(t)	h(t)
Weibull	$e^{-\lambda t^p}$	$\lambda p t^{p-1}$

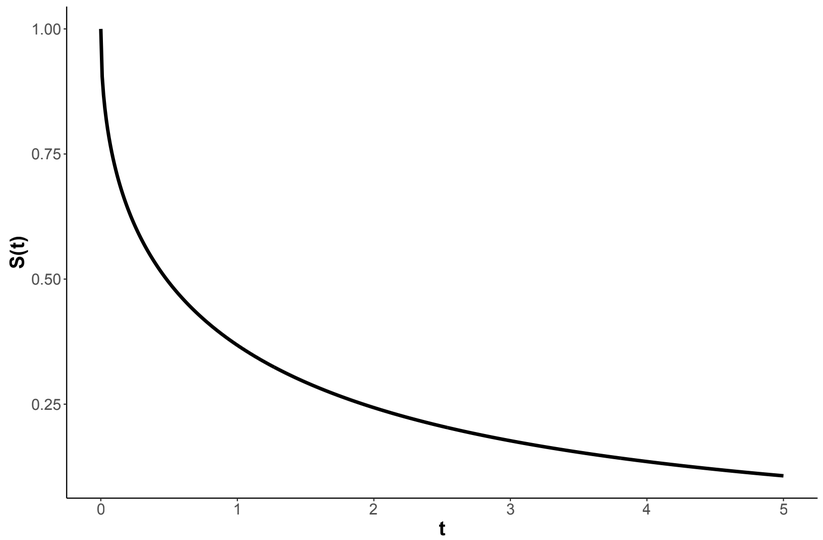
(λ is reparameterized for regression in terms of predictor variables)

github.com/nathansam/igmm-survival

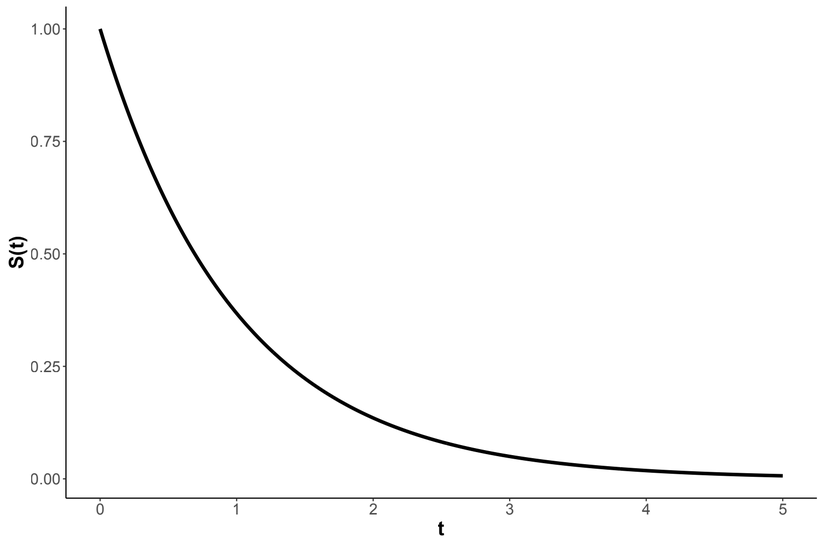
Accelerated failure time models

Weibull survival function plot

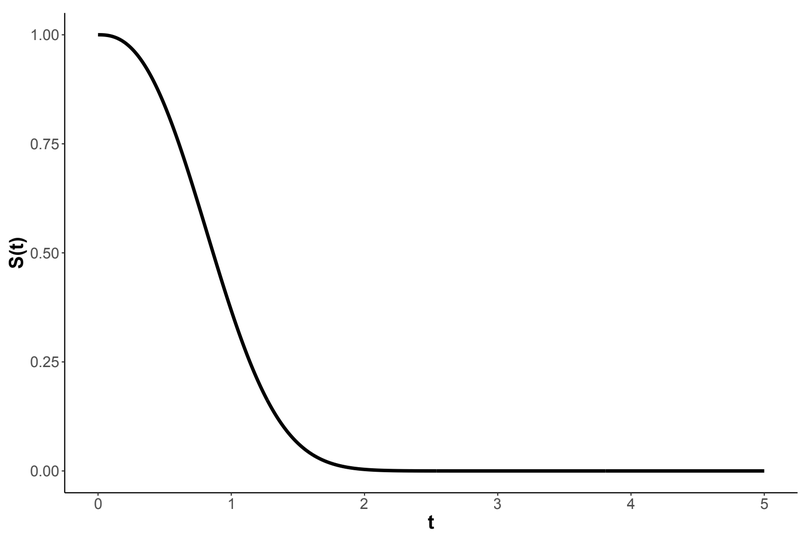
P = 0.5



P = 1



P = 2.5

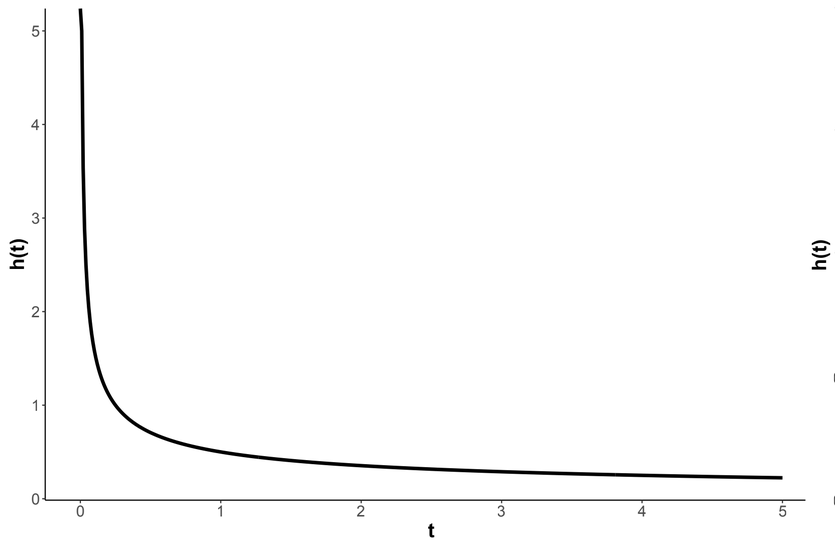


github.com/nathansam/igmm-survival

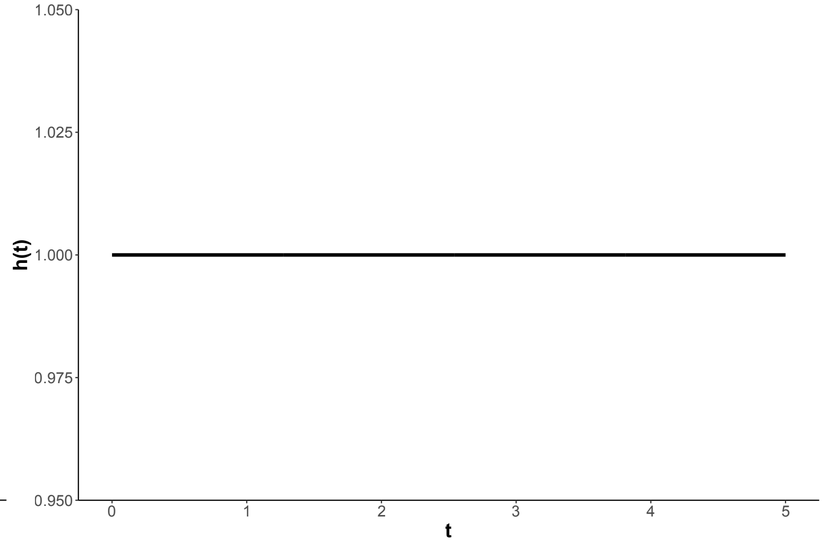
Accelerated failure time models

Weibull hazard function plot

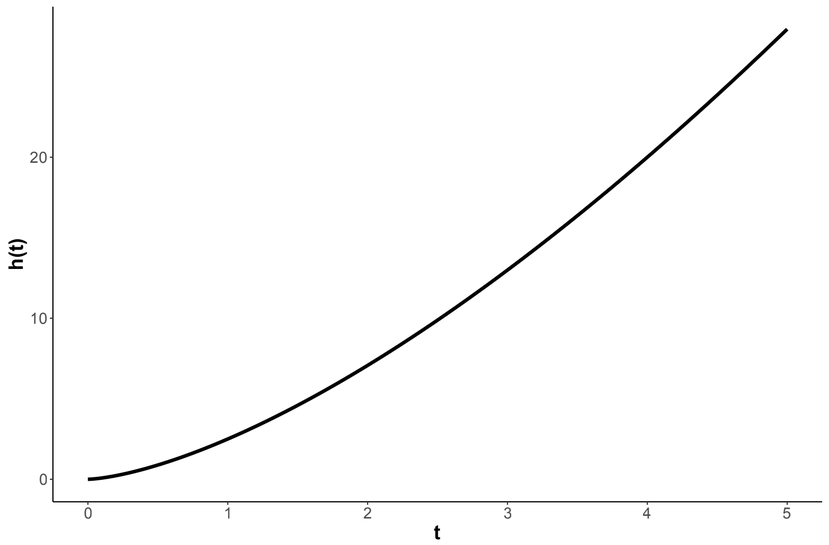
P = 0.5



P = 1



P = 2.5



github.com/nathansam/igmm-survival

Accelerated failure time models

The AFT assumption

- The underlying assumption for AFT models is that the effect of covariates is multiplicative with respect to survival time.
- $\log\left(T_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \epsilon_i$
- This also means the key measure of association for an AFT is an acceleration factor (instead of a hazard ratio)
- Although the Weibull PH and Weibull AFT models have different underlying assumptions, they are actually the same model (just parameterised differently)

github.com/nathansam/igmm-survival

Competing risks survival analysis

github.com/nathansam/igmm-survival

Competing risks survival analysis

Introduction to Competing Risks

- In the real world, there are many events which prevent another event from occurring such as:
    - Death due to disease of interest & death unrelated to disease
    - Hospital discharge & death
    - Ulcerative colitis flare & pan-proctocolectomy (removal of the colon, rectum and anus)

Competing risks survival analysis

Cause-specific approach

- This approach is the most common in competing risks, and involves performing survival analysis for each event type separately (the other event types are treated as censored)
- Kaplan-Meier-based survival curves have “questionable” interpretations in the context of competing risks
- An alternative to KM-based curves are conditional probability curves which provides a risk probability conditional on an individual not experiencing any of the other competing risks by time t.
- We assume the competing risks are independent (but we cannot verify this with observed data).

github.com/nathansam/igmm-survival

Competing risks survival analysis

How do we handle the independence assumption?

- Decide assumption is valid by clinical/ biological arguments
- Include common risk factors (e.g smoking for cardiovascular disease and cancer)
- Via a sensitivity analysis which considers “worst-case” violations

github.com/nathansam/igmm-survival

Joint modelling of time-to-event and longitudinal outcomes

github.com/nathansam/igmm-survival

Joint modelling of time-to-event and longitudinal outcomes

Introduction to JMs 1

- Up to now, all of our models have used covariates measured at baseline
- What if we wish to incorporate measurements which have been recorded across follow-up? For example:
    - C-reactive protein, CD4 counts, forced expiratory volume (FEV), faecal calprotectin, patient-reported stress
- Time-dependent Cox assumes longitudinal measurements are constant between measurements. Problematic.
- Instead we should model the longitudinal process (usually via a linear mixed effects) and the survival process (usually via Cox proportional hazards)

github.com/nathansam/igmm-survival

Joint modelling of time-to-event and longitudinal outcomes

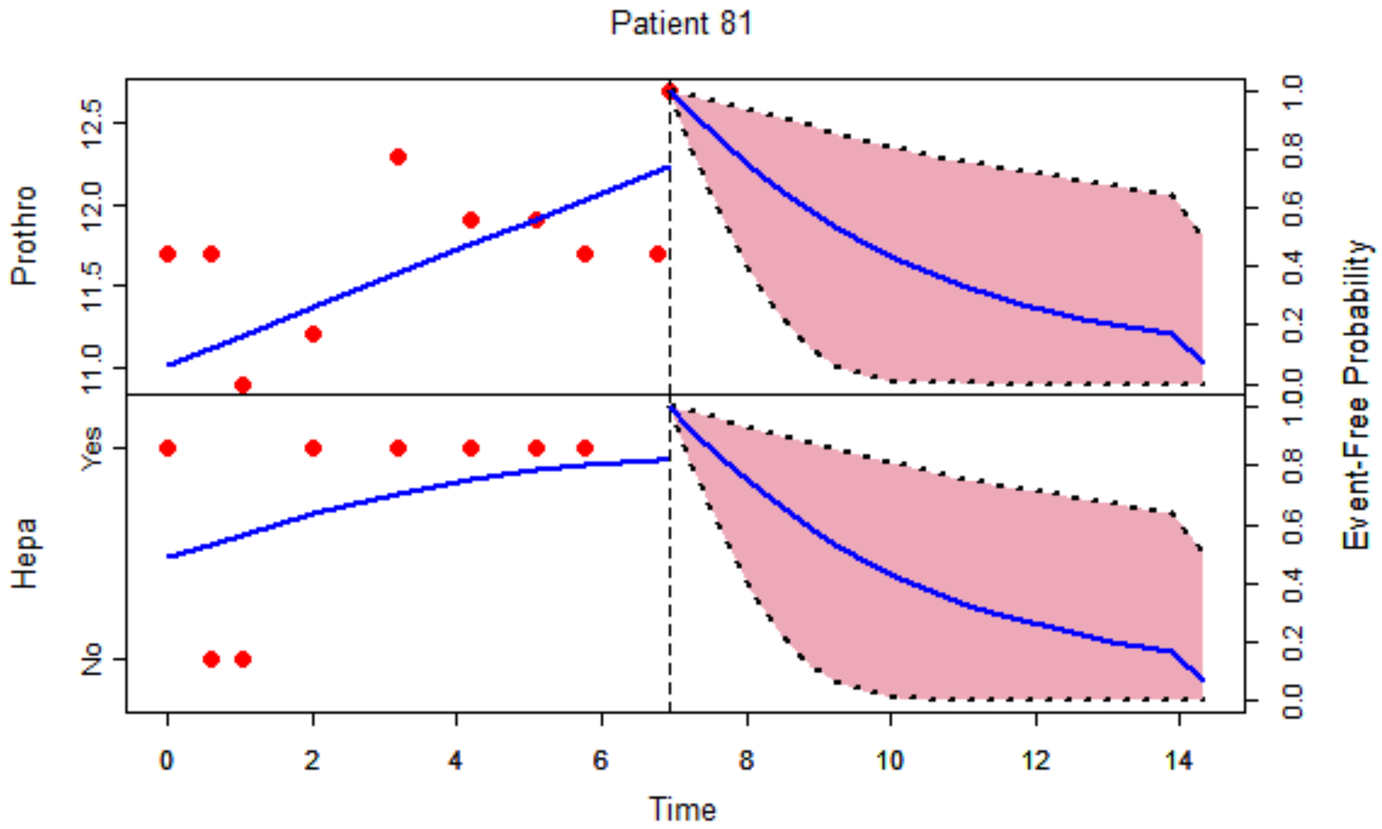
Introduction to JMs 2

- Modelling these two processes separately leads to biased estimates.
- Instead, we (typically) assume the survival and time-to-event models are interdependent via some shared random effects specific to the individual.
- We call the overall model a joint model (JM), the survival and longitudinal models submodels.
- Multivariate JMs allow for multiple longitudinal variables
- Being able to account for individual variability means we can characterise individual disease progression. This naturally has applicability for precision medicine.

github.com/nathansam/igmm-survival

Joint modelling of time-to-event and longitudinal outcomes

JM plot

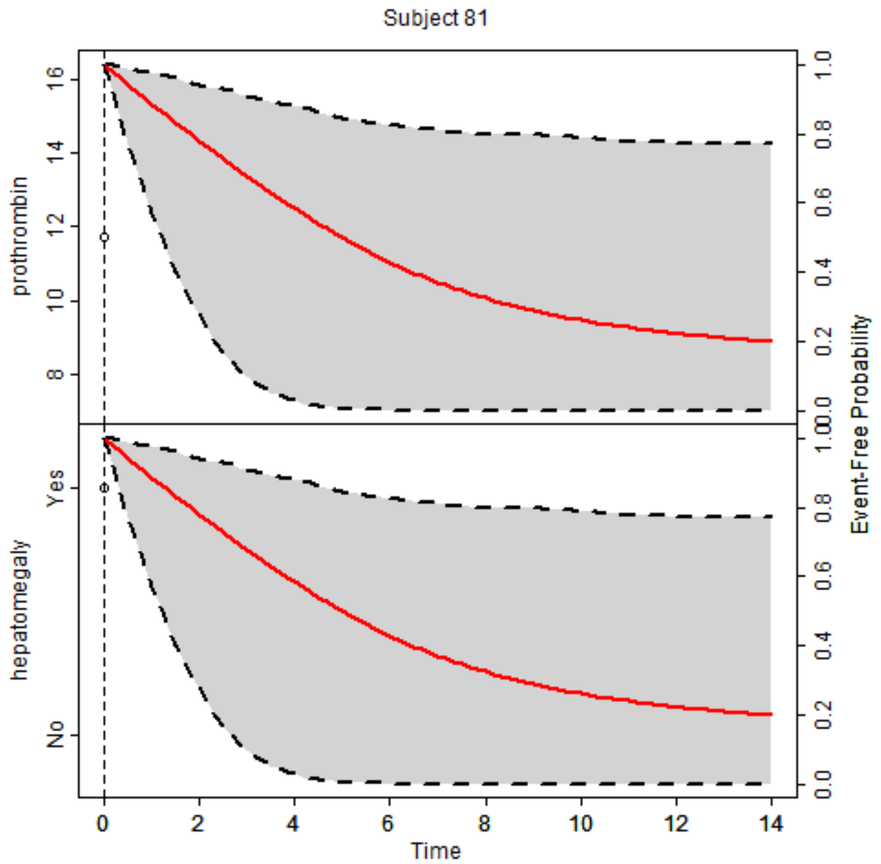


Rizopoulos, D. “The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC”. Journal of Statistical Software, **2016**, 72, 1-45

github.com/nathansam/igmm-survival

Joint modelling of time-to-event and longitudinal outcomes

Dynamic JM plot



Rizopoulos, D. “The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC”. Journal of Statistical Software, **2016**, 72, 1-45

github.com/nathansam/igmm-survival

Acknowledgements

Vallejos Group

Catalina Vallejos

James Liley

Karla Monterrubio-Gomez

Alan O'Callaghan

ChantrioInt-Andreas Kapourani

Evgenii Lobzaev

Christos Maniatis

Andrew Papanastasiou

Rachel Jackson

Lees Group / PREdiCCt

Charlie Lees

Gareth Jones

Spyros Siakavellas

Nik Plevris

Laura Lucaciu

Phil Jenkinson

Lisa Derr

Lauren Murdoch

github.com/nathansam/igmm-survival