A flexible approach to time-to-event data analysis using case-base sampling

Jesse Islam

July 11, 2019

Motivating example

- Meet Justin

Motivating example

- Meet Justin
    - Age: 56

Motivating example

- Meet Justin
    - Age: 56
    - Worried about his Prostate

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Justin's two year risk for death by Prostate Cancer?

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis
    - Cox Regression

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis
    - Cox Regression
- When we want the absolute risk:

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis
    - Cox Regression
- When we want the absolute risk:
    - Parametric models

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis
    - Cox Regression
- When we want the absolute risk:
    - Parametric models
    - Breslow estimator

Motivations for a new method

- Julien and Hanley (2008) found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages.

Motivations for a new method

- Julien and Hanley (2008) found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages.
- They believe that this is due to the Cumulative incidence curves (or survival curves) being stepwise rather than smooth, reducing interpretability.

Motivations for a new method

- Julien and Hanley (2008) found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages.
- They believe that this is due to the Cumulative incidence curves (or survival curves) being stepwise rather than smooth, reducing interpretability.
- Easily model non proportional hazards

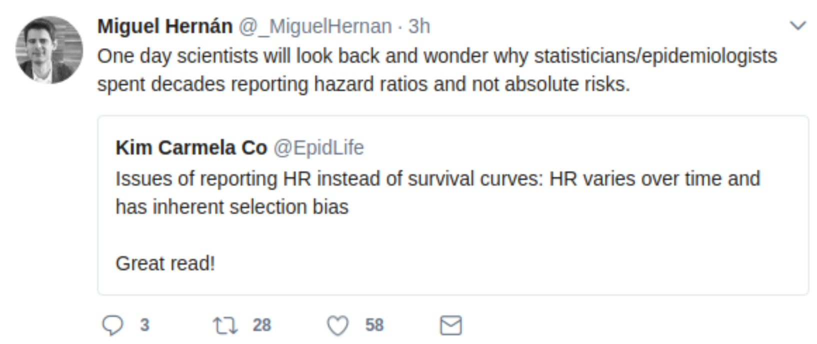
Motivations for a new method

- Julien and Hanley (2008) found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages.
- They believe that this is due to the Cumulative incidence curves (or survival curves) being stepwise rather than smooth, reducing interpretability.
- Easily model non proportional hazards
- Flexible fits

Motivations for a new method

- Julien and Hanley (2008) found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages.
- They believe that this is due to the Cumulative incidence curves (or survival curves) being stepwise rather than smooth, reducing interpretability.
- Easily model non proportional hazards
- Flexible fits
- A streamlined approach for reaching a **smooth absolute risk** curve

Dr. Hernan's perspective

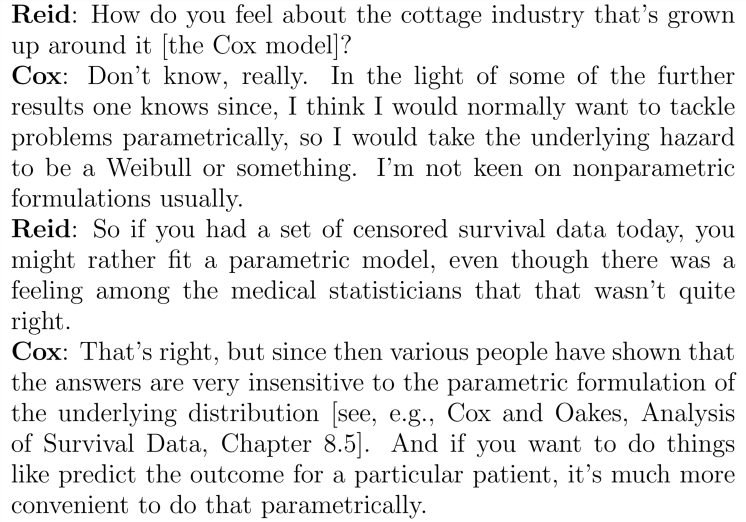


Miguel Hernán @_MiguelHernan · 3h

One day scientists will look back and wonder why statisticians/epidemiologists spent decades reporting hazard ratios and not absolute risks.

> **Kim Carmela Co** @EpidLife
>
> Issues of reporting HR instead of survival curves: HR varies over time and has inherent selection bias
>
> Great read!

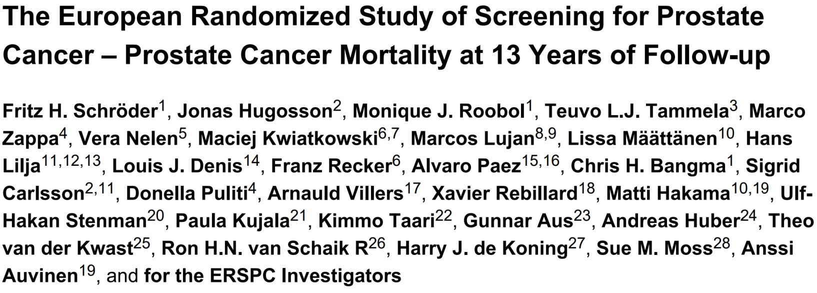
Dr. Cox's perspective



Reid: How do you feel about the cottage industry that's grown up around it [the Cox model]?
Cox: Don't know, really. In the light of some of the further results one knows since, I think I would normally want to tackle problems parametrically, so I would take the underlying hazard to be a Weibull or something. I'm not keen on nonparametric formulations usually.
Reid: So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right.
Cox: That's right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution [see, e.g., Cox and Oakes, Analysis of Survival Data, Chapter 8.5]. And if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically.

European Randomized Study of Prostate Cancer Screening (ERSPC) Data

- ~150 000 men ages 55-69



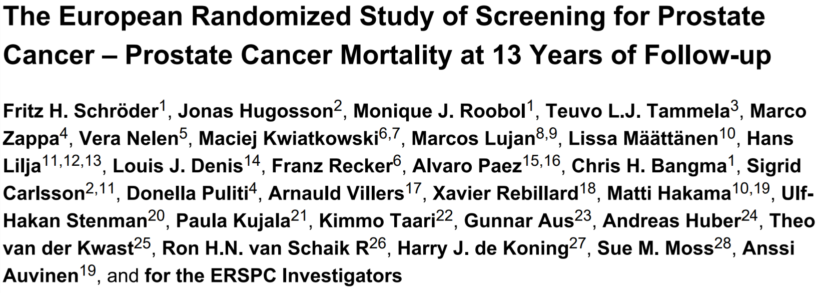
The European Randomized Study of Screening for Prostate Cancer – Prostate Cancer Mortality at 13 Years of Follow-up

Fritz H. Schröder[1], **Jonas Hugosson**[2], **Monique J. Roobol**[1], **Teuvo L.J. Tammela**[3], **Marco Zappa**[4], **Vera Nelen**[5], **Maciej Kwiatkowski**[6,7], **Marcos Lujan**[8,9], **Lissa Määttänen**[10], **Hans Lilja**[11,12,13], **Louis J. Denis**[14], **Franz Recker**[6], **Alvaro Paez**[15,16], **Chris H. Bangma**[1], **Sigrid Carlsson**[2,11], **Donella Puliti**[4], **Arnauld Villers**[17], **Xavier Rebillard**[18], **Matti Hakama**[10,19], **Ulf-Hakan Stenman**[20], **Paula Kujala**[21], **Kimmo Taari**[22], **Gunnar Aus**[23], **Andreas Huber**[24], **Theo van der Kwast**[25], **Ron H.N. van Schaik R**[26], **Harry J. de Koning**[27], **Sue M. Moss**[28], **Anssi Auvinen**[19], and **for the ERSPC Investigators**

Schroder FH, et al., for the ERSPC Investigators.Screening and Prostate-Cancer Mortality in a Randomized European Study. N Engl J Med 2009;360:1320-8.

European Randomized Study of Prostate Cancer Screening (ERSPC) Data

- ~150 000 men ages 55-69
- First start: 1991



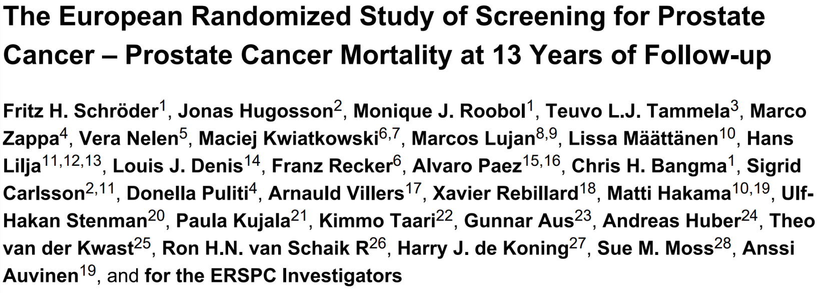
The European Randomized Study of Screening for Prostate Cancer – Prostate Cancer Mortality at 13 Years of Follow-up

Fritz H. Schröder[1], Jonas Hugosson[2], Monique J. Roobol[1], Teuvo L.J. Tammela[3], Marco Zappa[4], Vera Nelen[5], Maciej Kwiatkowski[6,7], Marcos Lujan[8,9], Lissa Määttänen[10], Hans Lilja[11,12,13], Louis J. Denis[14], Franz Recker[6], Alvaro Paez[15,16], Chris H. Bangma[1], Sigrid Carlsson[2,11], Donella Puliti[4], Arnauld Villers[17], Xavier Rebillard[18], Matti Hakama[10,19], Ulf-Hakan Stenman[20], Paula Kujala[21], Kimmo Taari[22], Gunnar Aus[23], Andreas Huber[24], Theo van der Kwast[25], Ron H.N. van Schaik R[26], Harry J. de Koning[27], Sue M. Moss[28], Anssi Auvinen[19], and for the ERSPC Investigators

Schroder FH, et al., for the ERSPC Investigators.Screening and Prostate-Cancer Mortality in a Randomized European Study. N Engl J Med 2009;360:1320-8.

European Randomized Study of Prostate Cancer Screening (ERSPC) Data

- ~150 000 men ages 55-69
- First start: 1991
- End: 2006



The European Randomized Study of Screening for Prostate Cancer – Prostate Cancer Mortality at 13 Years of Follow-up

Fritz H. Schröder[1], **Jonas Hugosson**[2], **Monique J. Roobol**[1], **Teuvo L.J. Tammela**[3], **Marco Zappa**[4], **Vera Nelen**[5], **Maciej Kwiatkowski**[6,7], **Marcos Lujan**[8,9], **Lissa Määttänen**[10], **Hans Lilja**[11,12,13], **Louis J. Denis**[14], **Franz Recker**[6], **Alvaro Paez**[15,16], **Chris H. Bangma**[1], **Sigrid Carlsson**[2,11], **Donella Puliti**[4], **Arnauld Villers**[17], **Xavier Rebillard**[18], **Matti Hakama**[10,19], **Ulf-Hakan Stenman**[20], **Paula Kujala**[21], **Kimmo Taari**[22], **Gunnar Aus**[23], **Andreas Huber**[24], **Theo van der Kwast**[25], **Ron H.N. van Schaik R**[26], **Harry J. de Koning**[27], **Sue M. Moss**[28], **Anssi Auvinen**[19], and **for the ERSPC Investigators**

Schroder FH, et al., for the ERSPC Investigators.Screening and Prostate-Cancer Mortality in a Randomized European Study. N Engl J Med 2009;360:1320-8.

ERSPC Data

```r
head(casebase::ERSPC)
```

PatientID	ScrArm	Follow.Up.Time	DeadOfPrCa
1	1	0.003	0
2	0	1.038	1
3	1	7.966	1
4	0	11.975	1
5	1	14.910	0

Recall

- Justin wants to know his two year risk for prostate cancer.

Recall

- Justin wants to know his two year risk for prostate cancer.
- As Justin was not part of the study, we will consider him part of the control group where no screening occured

Recall

- Justin wants to know his two year risk for prostate cancer.
- As Justin was not part of the study, we will consider him part of the control group where no screening occured
- **We will determine Justin's absolute risk using CaseBase!**

Casebase Overview

1. Clever sampling.

Casebase Overview

1. Clever sampling.
2. Indirectly deals with censoring.

Casebase Overview

1. Clever sampling.
2. Indirectly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

Casebase Overview

1. Clever sampling.
2. Indirectly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

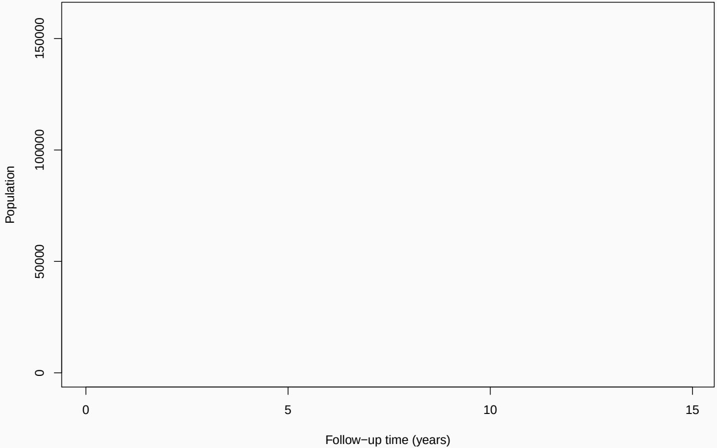
- Casebase is parametric, and allows different parametric fits by incorporation of the time component.

Casebase Overview

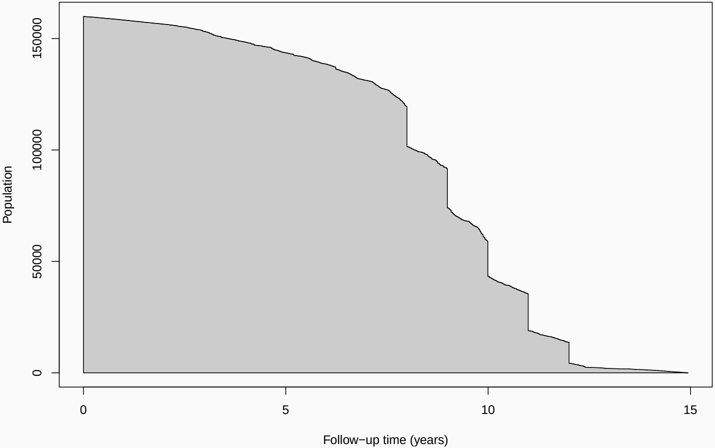
1. Clever sampling.
2. Indirectly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

- Casebase is parametric, and allows different parametric fits by incorporation of the time component.
- Package contains an implementation for generating *population-time* plots.

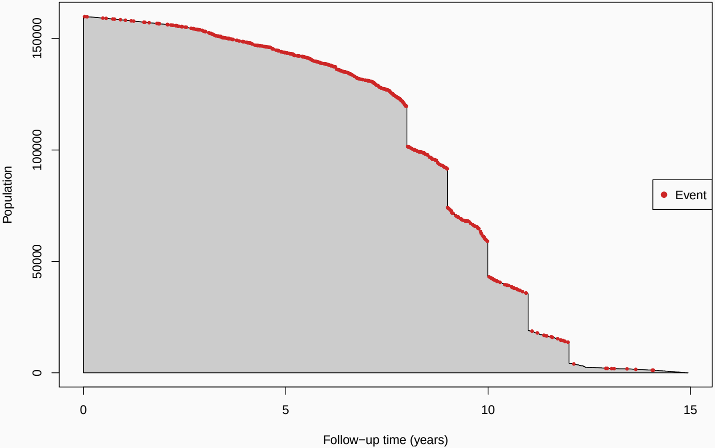
Casebase: Sampling



Casebase: Sampling

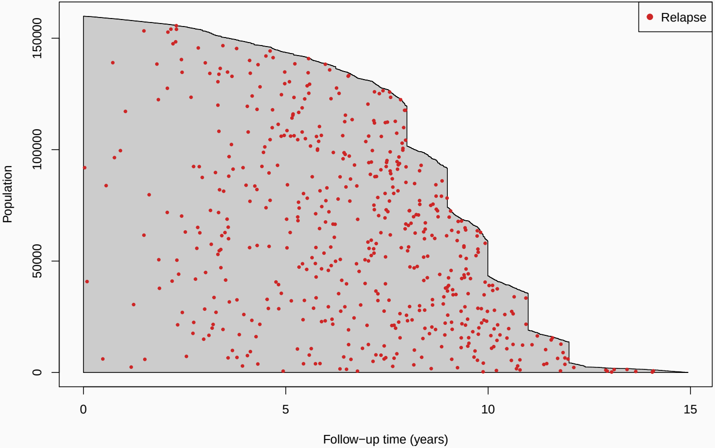


Casebase: Sampling

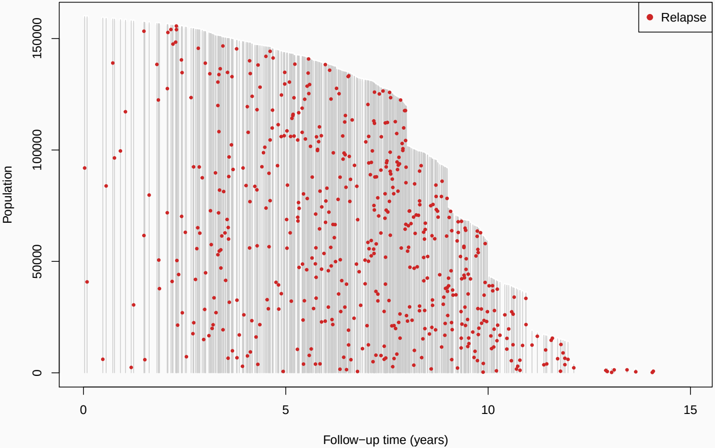


Casebase: Sampling

```
casebase::popTime(Data,Event,Time)
```



Casebase: Sampling



Casebase: Parametric families

- We can now fit models of the form:

$$log(h(t; \alpha, \beta)) = g(t; \alpha) + \beta X$$

Casebase: Parametric families

- We can now fit models of the form:

$$log(h(t;\alpha,\beta)) = g(t;\alpha) + \beta X$$

- By changing the function $g(t;\alpha)$, we can model different parametric families easily:

Casebase: Parametric models

Exponential: $g(t;\alpha)$ is equal to a constant

```
casebase::fitSmoothHazard(status ~ X)
```

Gompertz: $g(t;\alpha) = \alpha t$

```
casebase::fitSmoothHazard(status ~ time + X)
```

Weibull: $g(t;\alpha) = \alpha log(t)$

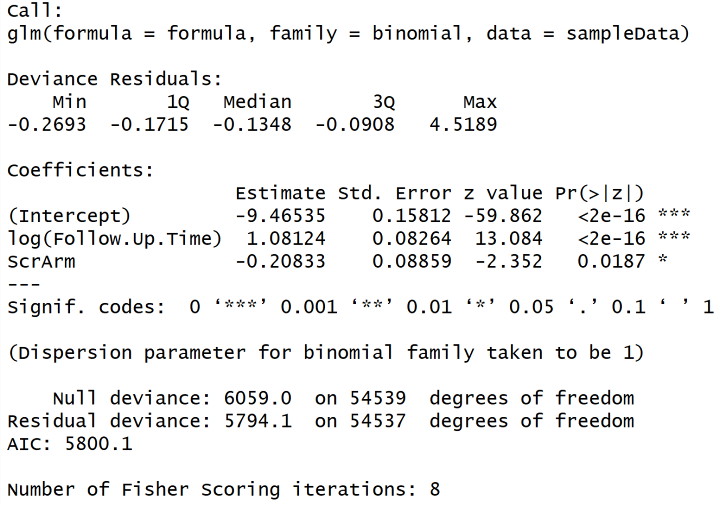
```
casebase::fitSmoothHazard(status ~ log(time) + X)
```

Casebase: Semi-Parametric models

Splines: $g(t;\alpha) = \alpha\ bs(t)$ casebase::fitSmoothHazard(status $\sim$ bs(time) $+$ X)

Prostate cancer hazard ratio

```
casebase::fitSmoothHazard(DeadOfPrCa ~ log(Follow.Up.Time)
                                       ScrArm, data=ERSPC, ratio = 100)
```



```
Call:
glm(formula = formula, family = binomial, data = sampleData)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.2693  -0.1715  -0.1348  -0.0908   4.5189  

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -9.46535    0.15812 -59.862   <2e-16 ***
log(Follow.Up.Time)  1.08124    0.08264  13.084   <2e-16 ***
ScrArm              -0.20833    0.08859  -2.352   0.0187 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6059.0  on 54539  degrees of freedom
Residual deviance: 5794.1  on 54537  degrees of freedom
AIC: 5800.1

Number of Fisher Scoring iterations: 8
```

ERSPC Hazard comparison

Model	Hazard Ratio	Std.Error
Cox	0.801	1.092
Gompertz	0.784	1.093
Exponential	0.809	1.092
Weibull	0.812	1.093
Splines	0.813	1.093

Absolute Risk

- We have a bunch of different parametric Hazard models now.

Absolute Risk

- We have a bunch of different parametric Hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x, t) = 1 - e^{-\int_0^t h(x,u)du}$$

Absolute Risk

- We have a bunch of different parametric Hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)

Absolute Risk

- We have a bunch of different parametric Hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)
- *h(x,u)*= Hazard Ratio

Absolute Risk

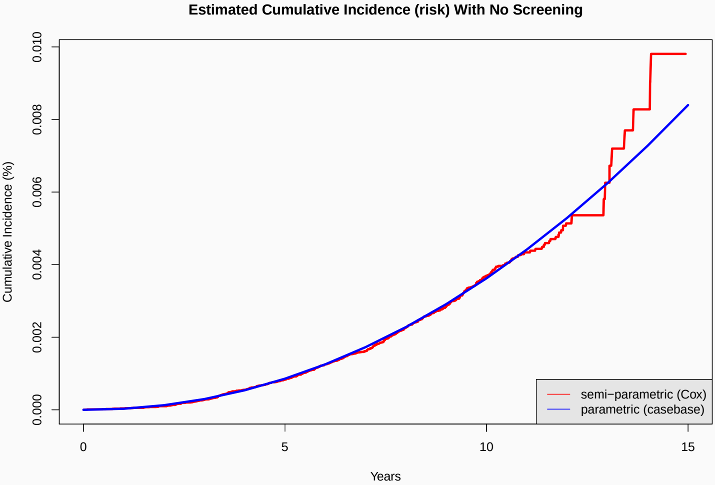
- We have a bunch of different parametric Hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)
- *h(x,u)*= Hazard Ratio
- Lets use the weibull hazard

Casebase: Absolute Risk comparison

```
casebase::absoluteRisk(Hazard, time, newdata)
```



Competing Risks

- Current methods:

Competing Risks

- Current methods:
    - Fine-Gray

Competing Risks

- Current methods:
  - Fine-Gray
  - Kaplan-Meier

Competing Risks

- Current methods:
  - Fine-Gray
  - Kaplan-Meier
- Proposed method:

Competing Risks

- Current methods:
  - Fine-Gray
  - Kaplan-Meier
- Proposed method:
  - Case-Base

Competing Risks: Data

- Two diseases:

```r
head(casebase::bmtcrr)
```

D	Status	ftime
ALL	2	0.67
AML	1	9.50
ALL	0	131.77
ALL	2	24.03

Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007

Competing Risks: Data

- Two diseases:
    - Lymphoblastic leukemia (ALL)

```r
head(casebase::bmtcrr)
```

D	Status	ftime
ALL	2	0.67
AML	1	9.50
ALL	0	131.77
ALL	2	24.03

Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007

Competing Risks: Data

- Two diseases:
  - Lymphoblastic leukemia (ALL)
  - Myeloblastic leukemia (AML)

```r
head(casebase::bmtcrr)
```

D	Status	ftime
ALL	2	0.67
AML	1	9.50
ALL	0	131.77
ALL	2	24.03

Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007

Competing Risks: Data

- Two diseases:
    - Lymphoblastic leukemia (ALL)
    - Myeloblastic leukemia (AML)
- Contains a competing event.

```r
head(casebase::bmtcrr)
```

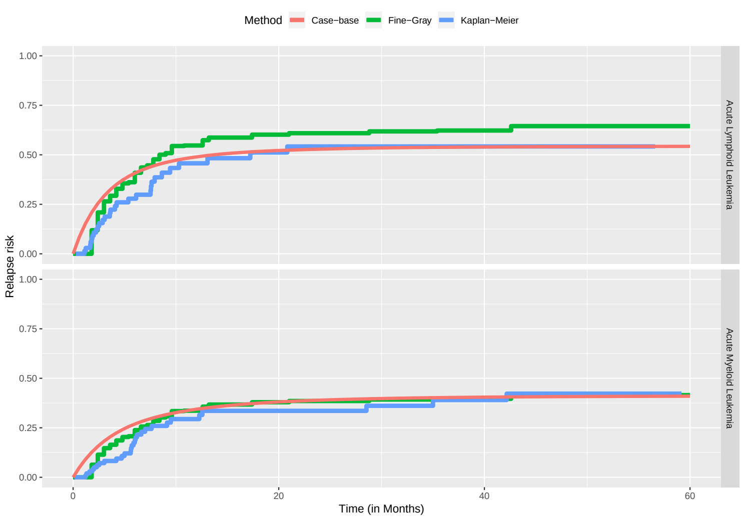
D	Status	ftime
ALL	2	0.67
AML	1	9.50
ALL	0	131.77
ALL	2	24.03

Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007

Competing Risks: Absolute Risk

```
model_cb <- casebase::fitSmoothHazard(Status ~ ftime
                                      + ... , data =
                                      bmtcrr)
risk_cb <- absoluteRisk(model_cb, Time, Newdata)
```

Competing Risks: Absolute Risk



Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
    - Population-Time plots

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
    - Population-Time plots
    - Hazard functions

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
  - Population-Time plots
  - Hazard functions
  - Absolute Risk

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
    - Population-Time plots
    - Hazard functions
    - Absolute Risk
    - Casebase can deal with competing risks.

References

http://sahirbhatnagar.com/casebase/ Math paper Hanley paper Max Presentation slides Olli presentation slides data reference data reference