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

Jesse Islam

July 11, 2019

Motivating example

- Meet Justin

Popular methods in time-to-event analysis

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56

Popular methods in time-to-event analysis

Motivations for a new method

Motivating example

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

Popular methods in time-to-event analysis

Motivations for a new method

Motivating example

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

Popular methods in time-to-event analysis

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - Cox Regression

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - Cox Regression
- When we want the absolute risk:

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - Cox Regression
- When we want the absolute risk:
  - Parametric models

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - Cox Regression
- When we want the absolute risk:
  - Parametric models
  - Breslow estimator

Motivations for a new method

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - 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

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - 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

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - 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

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - 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

Motivating example

- Meet Justin
  - Age: 56
  - Worried about his Prostate
  - What is Carol'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
  - 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

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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**

ERSPC Data

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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**

ERSPC Data

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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**

ERSPC Data

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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**

ERSPC Data

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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**

ERSPC Data

Data on the men in the European Randomized Study of Prostate Cancer Screening (ERSPC)

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

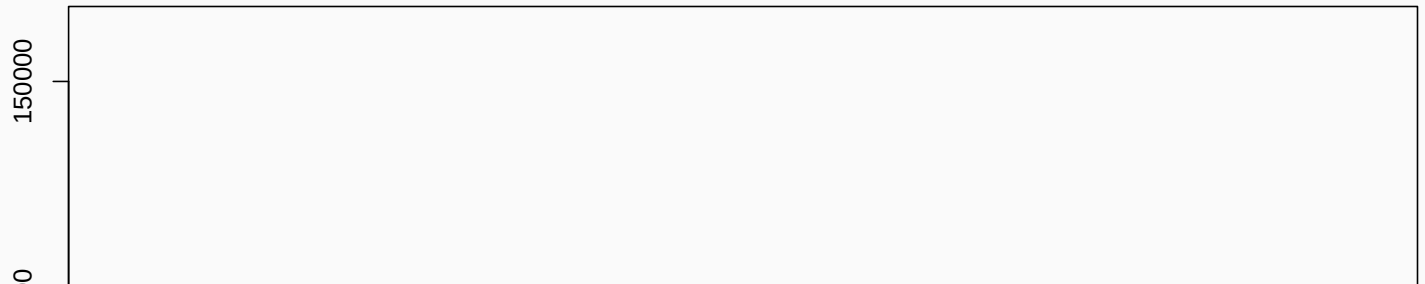
Fritz H. Schröder[1], **Jonas Hugosson**[2], **Monique J. Roobol**[1], **Teuvo L.J. Tammela**[3], **Marco Zappa**[4], **Vera Nelen**[5], **Maciej Kwiatkowski**[6,7], **Marcos Lujan**[8,9], **Lissa Määttänen**[10], **Hans Lilja**[11,12,13], **Louis J. Denis**[14], **Franz Recker**[6], **Alvaro Paez**[15,16], **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**

ERSPC Data

Casebase Overview

1. Clever sampling.

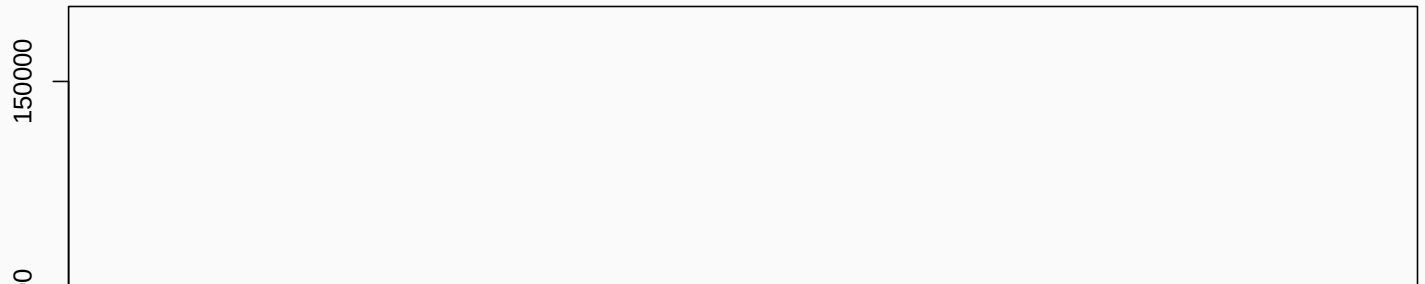
Casebase: Sampling



Casebase Overview

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

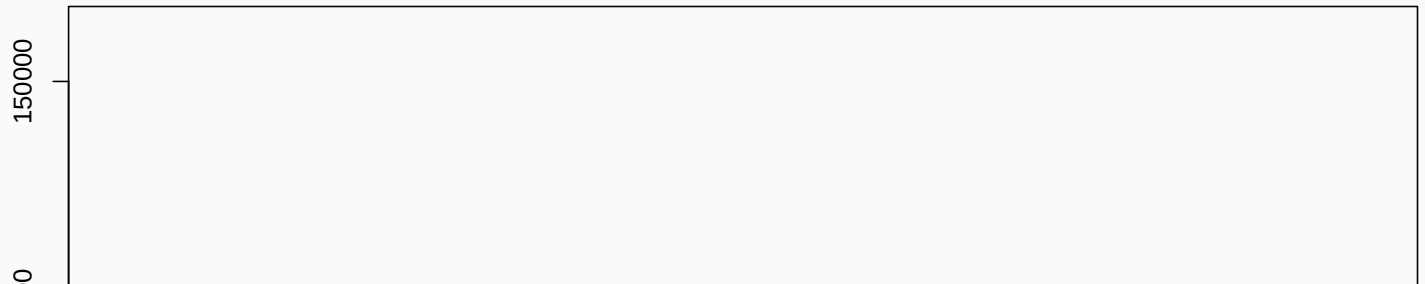
Casebase: Sampling



Casebase Overview

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

Casebase: Sampling

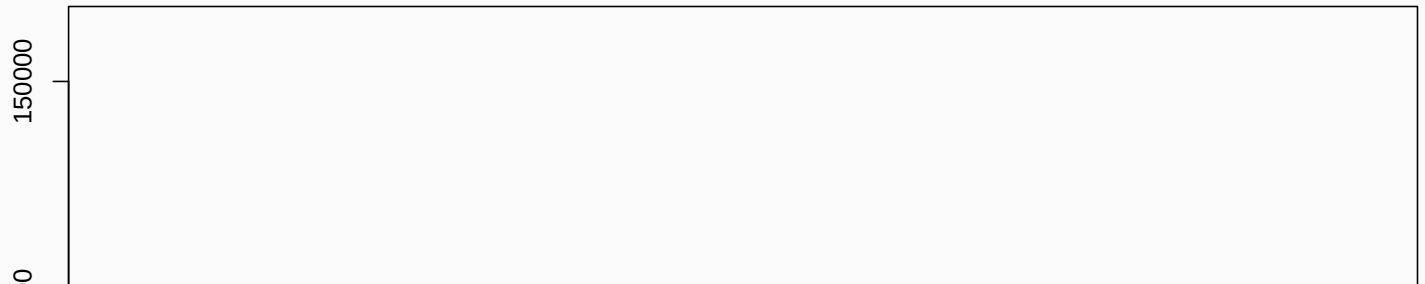


Casebase Overview

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: Sampling

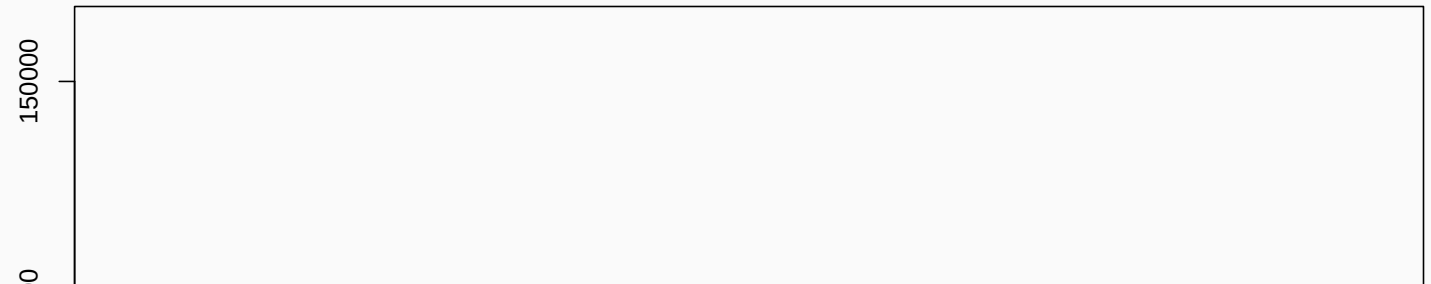


Casebase Overview

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: Parametric families

- We can now fit models of the form:

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

Casebase: Parametric models

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

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

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

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

Casebase: Hazard fits

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 ~ var)
```

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

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

ERSPC Hazard

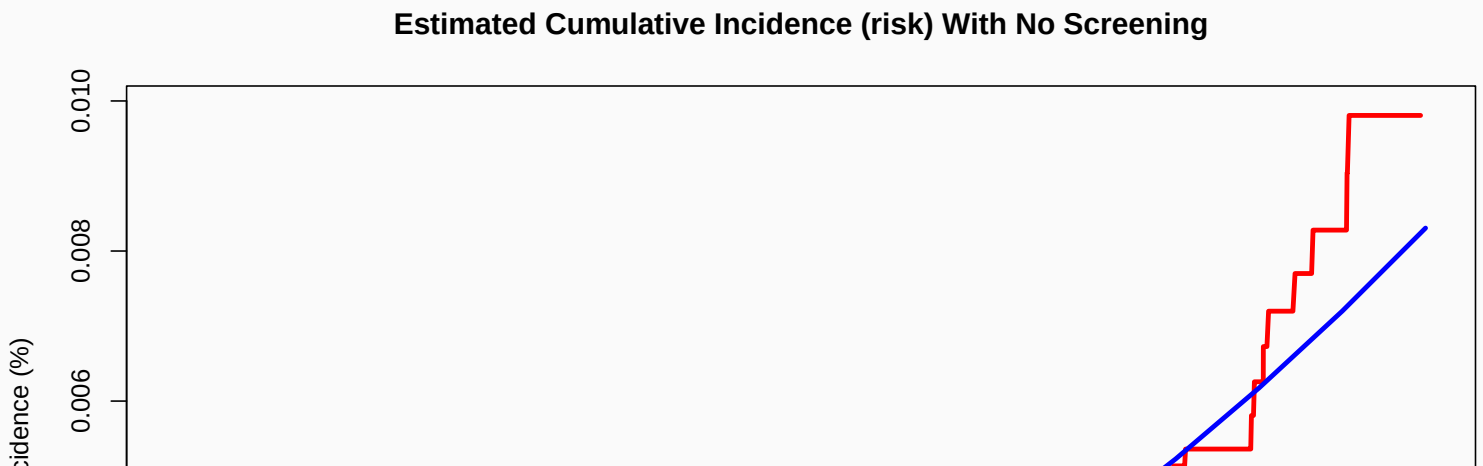
```
wModel<-fitSmoothHazard(DeadOfPrCa ~
log(Follow.Up.Time) +
ScrArm, data=ERSPC, ratio = 100)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2678  -0.1715  -0.1347  -0.0908   4.5127

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.44715    0.15750 -59.984   <2e-16 ***
log(Follow.Up.Time)  1.07406    0.08237  13.039   <2e-16 ***
ScrArm1            -0.22362    0.08859  -2.524   0.0116 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Absolute Risk

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

Casebase: Absolute Risk comparison

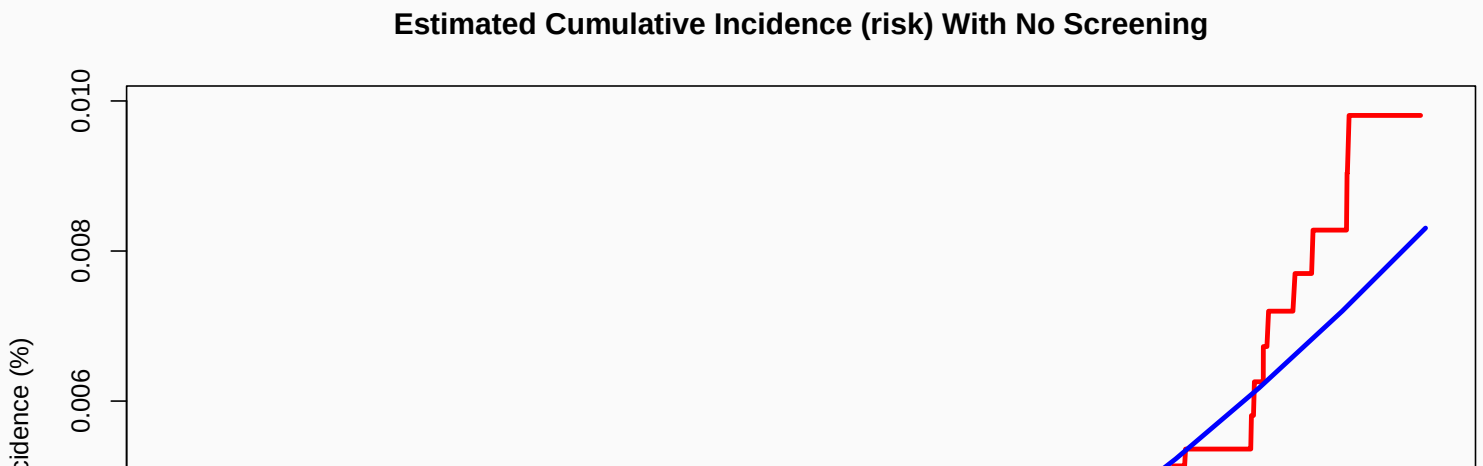


Absolute Risk

- 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:

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

Casebase: Absolute Risk comparison



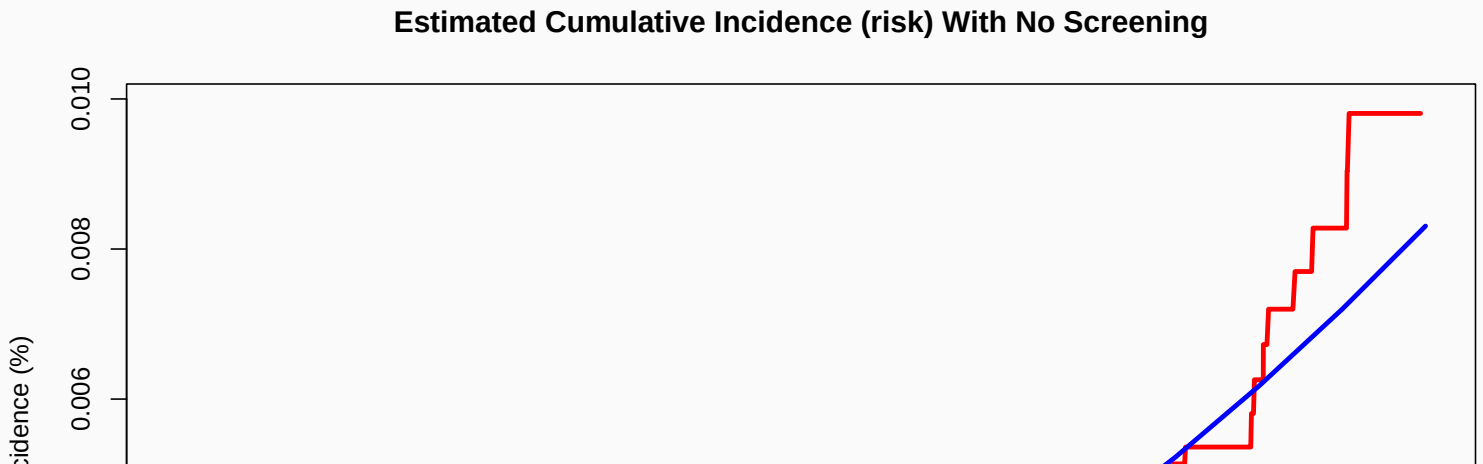
Absolute Risk

- 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:

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

- Lets use the weibull hazard

Casebase: Absolute Risk comparison



Competing Risks

- Current methods:

Competing Risks: Data

D	Status	ftime
ALL	2	0.67

Competing Risks

- Current methods:
- Fine-Gray

Competing Risks: Data

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

- Two diseases:

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

- Two diseases:
- Lymphoblastic leukemia (ALL)

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

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

D	Status	ftime
ALL	2	0.67

Competing Risks

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

Competing Risks: Data

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

D	Status	ftime
ALL	2	0.67

Summary

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

References

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

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:

References

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

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

References

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

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

References

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

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

References

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

Summary

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

cumulative incidence function for the Cox model

plot(coxRisk*time*, *coxRisk*cumhaz, type="l", xlab = "Years", ylab = "Cumulative Incidence (%)", fun = "event", xlim = c(0,15), conf.int = F, col = "red", main = sprintf("Estimated Cumulative Incidence (risk)"))

add casebase curve with legend

lines(wRisk[,1], wRisk[,2], type = “l”, col = “blue”)
legend(“bottomright”, legend = c(“semi-parametric (Cox)”, “parametric (casebase)”), col = c(“red”,“blue”), lty = c(1, 1), bg = “gray90”)