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

Jesse Islam
McGill University
July 11, 2019
Motivating example

- Meet Justin.

- Age: 56
- Worried about his prostate.
- What is Justin's two year risk of death due to prostate cancer?
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 of death due to prostate cancer?
Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.
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
Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [1]
Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [1]
- They believe the stepwise nature is the reason, as it reduces interpretability. [1]
Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [1]
- They believe the stepwise nature is the reason, as it reduces interpretability. [1]
- Want to easily model non-proportional hazards. [1]
Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [1]
- They believe the stepwise nature is the reason, as it reduces interpretability. [1]
- Want to easily model non-proportional hazards. [1]
- A streamlined approach for reaching a smooth absolute risk curve. [1]
**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. [4]
European Randomized Study of Prostate Cancer Screening (ERSPC) Data

- ~150,000 men ages 55-69. [4]
- Examined effects screening has on death due to prostate cancer. [4]

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

Fritz H. Schröder¹, Jonas Hugosson², Monique J. Roobol¹, Teuvo L.J. Tammela³, Marco Zappa⁴, Vera Nelen⁵, Maciej Kwiatkowski⁶,⁷, Marcos Lujan⁸,⁹, Lissa Määtänn¹⁰, Hans Lilja¹¹,¹²,¹³, Louis J. Denis¹⁴, Franz Recker⁶, Alvaro Paez¹⁵,¹⁶, Chris H. Bangma¹, Sigrid Carlsson²,¹¹, Donella Puliti⁴, Arnauld Villers¹⁷, Xavier Rebillard¹⁸, Matti Hakama¹⁰,¹⁹, Ulf-Hakan Stenman²⁰, Paula Kujala²¹, Kimmo Taari²², Gunnar Aus²³, Andreas Huber²⁴, Theo van der Kwast²⁵, Ron H.N. van Schaik R²⁶, Harry J. de Koning²⁷, Sue M. Moss²⁸, Anssi Auvinen¹⁹, and for the ERSPC Investigators
### ERSPC Data

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

<table>
<thead>
<tr>
<th>PatientID</th>
<th>ScrArm</th>
<th>Follow.Up.Time</th>
<th>DeadOfPrCa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.038</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7.966</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>11.975</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>14.910</td>
<td>0</td>
</tr>
</tbody>
</table>
Using the ERSPC dataset and casebase, we will determine Justin’s absolute risk for death by prostate cancer.
1. Clever sampling.
1. Clever sampling.
2. Allows a parametric fit using *logistic regression*.
1. Clever sampling.
2. Allows a parametric fit using *logistic regression*.

- Casebase is parametric, and allows different parametric fits by incorporation of the time component.
1. Clever sampling.
2. Allows a parametric fit using \textit{logistic regression}.

- Casebase is parametric, and allows different parametric fits by incorporation of the time component.
- Package contains an implementation for generating \textit{population-time} plots.
Casebase: Sampling [5]
Casebase: Sampling [5]

![Graph showing population and death by prostate cancer over follow-up time.](image)
casebase::popTime(Data, Event, Time)
Casebase: Sampling [3]
We can now fit models of the form: [1]

$$\log(h(t; \alpha, \beta)) = g(t; \alpha) + \beta X$$
We can now fit models of the form: [1]

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

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

```r
casebase::fitSmoothHazard(status ~ X1 + X2)
```

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

```r
casebase::fitSmoothHazard(status ~ time + X1 + X2)
```

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

```r
casebase::fitSmoothHazard(status ~ \log(time) + X1 + X2)
```
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Hazard Ratio</th>
<th>Std.Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox</td>
<td>0.801</td>
<td>1.092</td>
</tr>
<tr>
<td>Gompertz</td>
<td>0.802</td>
<td>1.093</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.810</td>
<td>1.092</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.797</td>
<td>1.093</td>
</tr>
</tbody>
</table>
Absolute Risk

- We have parametric hazard models now.

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

- CI(x, t) = Cumulative Incidence (Absolute Risk)
- \( h(x, u) \) = Hazard function
- Let's use the Weibull hazard.
- We have 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 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 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 function
Absolute Risk

- We have 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 function
- Let's use the Weibull hazard.
Casebase: Absolute Risk comparison

casebase::absoluteRisk(fit, time=2, covariate_profile)

Estimated Cumulative Incidence (risk) With No Screening

Cumulative Incidence (%)

semi-parametric (Cox)
parametric (casebase)
Summary

- Casebase sampling permits the use of GLMs and the tools associated with them
Casebase sampling 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
- Flexible fits through splines
- Casebase can deal with competing risks.
Casebase sampling permits the use of GLMs and the tools associated with them.

The casebase package contains tools to generate:

- Population-Time plots
Casebase sampling 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
- Flexible fits through splines.
Casebase sampling 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 sampling 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

Flexible fits through splines.
Casebase sampling 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

Flexible fits through splines.
Casebase can deal with competing risks.


Tutorial and Slides

Tutorial:
http://sahirbhatnagar.com/casebase/

Slides:
https://github.com/Jesse-Islam/UseR–CaseBase-Presentation

Questions?
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)
```

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>Status</th>
<th>ftim e</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>2</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>1</td>
<td>9.50</td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>0</td>
<td>131.77</td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>24.03</td>
<td></td>
</tr>
</tbody>
</table>
Competing Risks: Data

- Two diseases:
  - Acute Lymphoblastic Leukemia (ALL)

\texttt{head(casebase::bmtcrr)}

<table>
<thead>
<tr>
<th>D</th>
<th>Status</th>
<th>ftime</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td>AML</td>
<td>1</td>
<td>9.50</td>
</tr>
<tr>
<td>ALL</td>
<td>0</td>
<td>131.77</td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>24.03</td>
</tr>
</tbody>
</table>
Competing Risks: Data

- Two diseases:
  - Acute Lymphoblastic Leukemia (ALL)
  - Acute Myeloblastic Leukemia (AML)

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

<table>
<thead>
<tr>
<th>D</th>
<th>Status</th>
<th>ftmte</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td>AML</td>
<td>1</td>
<td>9.50</td>
</tr>
<tr>
<td>ALL</td>
<td>0</td>
<td>131.77</td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>24.03</td>
</tr>
</tbody>
</table>
Competing Risks: Data

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

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

<table>
<thead>
<tr>
<th>D</th>
<th>Status</th>
<th>ftime</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td>AML</td>
<td>1</td>
<td>9.50</td>
</tr>
<tr>
<td>ALL</td>
<td>0</td>
<td>131.77</td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>24.03</td>
</tr>
</tbody>
</table>
fit_cb <- casebase::fitSmoothHazard(Status ~ ftime
+ ... , data =
  bmtcrr)

risk_cb <- absoluteRisk(fit_cb, Time, Newdata)
Competing Risks: Absolute Risk

- Acute Lymphoid Leukemia
- Acute Myeloid Leukemia

Method: Case-base, Fine-Gray, Kaplan-Meier

Relapse risk over time (in months). The graph compares the methods for estimating competing risks in acute lymphoid and myeloid leukemias.