A novel approach to competing risks analysis using case-base sampling

Maxime Turgeon
June 10th, 2017

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health
Acknowledgements

This project is joint work with:

- Sahir Bhatnagar
- Olli Saarela (U. Toronto)
- Jim Hanley
Introduction
Motivation

• Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia

• "What is my 5-year risk of relapse?"

• \( P(\text{Time to event} < 5, \text{Relapse} | \text{Covariates}) \)

• "What about 1-year? 2-year?"

• A smooth absolute risk curve.
• Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- “What is my 5-year risk of relapse?”
Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- “What is my 5-year risk of relapse?”
  - $P(\text{Time to event} < 5, \text{Relapse} \mid \text{Covariates})$
Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- “What is my 5-year risk of relapse?”
  - $P(\text{Time to event} < 5, \text{Relapse} \mid \text{Covariates})$
- “What about 1-year? 2-year?”
Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- “What is my 5-year risk of relapse?”
  - $P(\text{Time to event} < 5, \text{Relapse} \mid \text{Covariates})$
- “What about 1-year? 2-year?”
  - A **smooth** absolute risk curve.
Current methods

• Proportional hazards hypothesis
  • Disease etiology
  • E.g. Cox regression.

• Proportional subdistribution hypothesis
  • Absolute risk
  • E.g. Fine-Gray model.
Current methods

- Proportional hazards hypothesis
Current methods

- Proportional hazards hypothesis
  - Disease etiology

- Proportional subdistribution hypothesis

- Absolute risk
  - E.g. Fine-Gray model.
Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
- Proportional subdistribution hypothesis
Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
- Proportional subdistribution hypothesis
  - Absolute risk
Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.

- Proportional subdistribution hypothesis
  - Absolute risk
  - E.g. Fine-Gray model.
We propose a simple approach to modeling directly the cause-specific hazards using (smooth) parametric families.

Our approach relies on Hanley & Miettinen's case-base sampling method [1].

Smooth hazards give rise to smooth absolute risk curves.

Our approach allows for a symmetric treatment of all time variables.

Finally, it also allows for hypothesis testing and variable selection.

This method is currently available in the R package casebase on CRAN. See also our website: http://sahirbhatnagar.com/casebase/
Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.

- Our approach relies on Hanley & Miettinen's case-base sampling method [1].

- Smooth hazards give rise to smooth absolute risk curves.

- Our approach allows for a **symmetric** treatment of all time variables.

- Finally, it also allows for **hypothesis testing** and **variable selection**.

This method is currently available in the R package **casebase** on CRAN. See also our website: http://sahirbhatnagar.com/casebase/
We propose a simple approach to modeling directly the cause-specific hazards using (smooth) parametric families. Our approach relies on Hanley & Miettinen’s case-base sampling method [1]. Smooth hazards give rise to smooth absolute risk curves. Our approach allows for a symmetric treatment of all time variables. Finally, it also allows for hypothesis testing and variable selection. This method is currently available in the R package casebase on CRAN. See also our website: http://sahirbhatnagar.com/casebase/
Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
  - Our approach relies on Hanley & Miettinen’s **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
Summary

• We propose a simple approach to modeling directly the cause-specific hazards using (smooth) parametric families.
  • Our approach relies on Hanley & Miettinen’s case-base sampling method [1].
• Smooth hazards give rise to smooth absolute risk curves.
• Our approach allows for a symmetric treatment of all time variables.

This method is currently available in the R package casebase on CRAN. See also our website: http://sahirbhatnagar.com/casebase/
Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
  - Our approach relies on Hanley & Miettinen’s **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
- Our approach allows for a **symmetric** treatment of all time variables.
- Finally, it also allows for **hypothesis testing** and **variable selection**.

This method is currently available in the R package **casebase** on CRAN. See also our website: [http://sahirbhatnagar.com/casebase/](http://sahirbhatnagar.com/casebase/).
Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
  - Our approach relies on Hanley & Miettinen’s **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
- Our approach allows for a **symmetric** treatment of all time variables.
- Finally, it also allows for **hypothesis testing** and **variable selection**.

This method is currently available in the R package `casebase` on CRAN.

See also our website: http://sahirbhatnagar.com/casebase/
Case-base sampling
Follow-up time (months)

Population

- Red circles: Relapse
- Blue diamonds: Competing event

6/19
Case-base sampling

- The unit of analysis is a person-moment.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
- The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.
- This framework can easily be used with time-varying covariates (e.g. time-varying exposure).
Case-base sampling

- The unit of analysis is a *person-moment*. 
Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinominal regression.
Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
  - The sampling process is taken into account using an offset term.
Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
  - The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.
Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
  - The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.
- This framework can easily be used with time-varying covariates (e.g. time-varying exposure).
Theoretical details
We make the following assumptions:
We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$. 
Assumptions

We make the following assumptions:

- For each event type \( j = 1, \ldots, m \), a non-homogeneous Poisson process with hazard \( \lambda_j(t) \).
  - At most one event type can occur.

- At most one event type can occur.
Assumptions

We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.
  - At most one event type can occur.
- Non-informative censoring.
We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.
  - At most one event type can occur.
- Non-informative censoring.
- Case-base sampling occurs following a non-homogenous Poisson process with hazard $\rho(t)$.
Each person-moment’s contribution to the likelihood is of the form:

\[ \prod_{j=1}^{m} \frac{\lambda_j(t)^{dN_j(t)}}{\rho(t) + \sum_{j=1}^{m} \lambda_j(t)}. \]
Each person-moment’s contribution to the likelihood is of the form:

$$\prod_{j=1}^{m} \frac{\lambda_j(t)^{dN_j(t)}}{\rho(t) + \sum_{j=1}^{m} \lambda_j(t)}.$$

This is reminiscent of a multinomial likelihood, with offset \(\log(1/\rho(t))\).
Likelihood

Main Theorem

The likelihood defined above has mean zero and is asymptotically normal. Implication: All the GLM machinery (e.g. deviance tests, information criteria, regularization) is available to us.
Main Theorem
The likelihood defined above has mean zero and is asymptotically normal.
Main Theorem
The likelihood defined above has mean zero and is asymptotically normal.

Implication: All the GLM machinery (e.g. deviance tests, information criteria, regularization) is available to us.
Parametric families

We can fit any model of the following form:

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

Different choices of the function $g$ lead to familiar parametric families:

- **Exponential**: $g$ is constant.
- **Gompertz**: $g(t; \alpha) = \alpha t$.
- **Weibull**: $g(t; \alpha) = \alpha \log t$. 
We can fit any model of the following form:

\[ \log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X. \]
We can fit any model of the following form:

\[ \log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X. \]

Different choices of the function \( g \) leads to familiar parametric families:
We can fit any model of the following form:

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

Different choices of the function $g$ leads to familiar parametric families:

- Exponential: $g$ is constant.
We can fit any model of the following form:

\[ \log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X. \]

Different choices of the function \( g \) leads to familiar parametric families:

- **Exponential**: \( g \) is constant.
- **Gompertz**: \( g(t; \alpha) = \alpha t \).
We can fit any model of the following form:

\[ \log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X. \]

Different choices of the function \( g \) leads to familiar parametric families:

- **Exponential**: \( g \) is constant.
- **Gompertz**: \( g(t; \alpha) = \alpha t. \)
- **Weibull**: \( g(t; \alpha) = \alpha \log t. \)
Simulation study
• We simulate 1000 datasets from an exponential and a Gompertz family.
Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate
Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate
- Random censoring
Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate
- Random censoring
- We compare case-base with a correctly specified family, case-base with splines, and Cox regression.
Simulation results

Exponential

Gompertz

<table>
<thead>
<tr>
<th>Method</th>
<th>Exponential</th>
<th>Gompertz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-base</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case-base/Splines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case-base</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case-base/Splines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Beta
Data analysis
## Data

<table>
<thead>
<tr>
<th>Variable description</th>
<th>Statistical summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>M=Male (87)</td>
</tr>
<tr>
<td></td>
<td>F=Female (72)</td>
</tr>
<tr>
<td>Disease</td>
<td>ALL (59)</td>
</tr>
<tr>
<td></td>
<td>AML (100)</td>
</tr>
<tr>
<td>Phase</td>
<td>CR1 (43)</td>
</tr>
<tr>
<td></td>
<td>CR2 (40)</td>
</tr>
<tr>
<td></td>
<td>CR3 (10)</td>
</tr>
<tr>
<td></td>
<td>Relapse (65)</td>
</tr>
<tr>
<td>Type of transplant</td>
<td>BM+PB (15)</td>
</tr>
<tr>
<td></td>
<td>PB (144)</td>
</tr>
<tr>
<td>Age of patient (years)</td>
<td>16–62</td>
</tr>
<tr>
<td></td>
<td>33 (IQR 19.5)</td>
</tr>
<tr>
<td>Failure time (months)</td>
<td>0.13–131.77</td>
</tr>
<tr>
<td></td>
<td>20.28 (30.78)</td>
</tr>
<tr>
<td>Status indicator</td>
<td>0=censored (40)</td>
</tr>
<tr>
<td></td>
<td>1=relapse (49)</td>
</tr>
<tr>
<td></td>
<td>2=competing event (70)</td>
</tr>
</tbody>
</table>
Absolute risk for female patient, median age, in relapse at transplant (stem cells from peripheral blood).
## Model fit

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case-base</th>
<th></th>
<th>Cox regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard ratio</td>
<td>95% CI</td>
<td>Hazard ratio</td>
<td>95% CI</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td>0.64</td>
<td>(0.35, 1.20)</td>
<td>0.75</td>
<td>(0.42, 1.35)</td>
</tr>
<tr>
<td><strong>Disease</strong></td>
<td>0.54</td>
<td>(0.27, 1.07)</td>
<td>0.63</td>
<td>(0.34, 1.19)</td>
</tr>
<tr>
<td><strong>Phase CR2</strong></td>
<td>1.00</td>
<td>(0.37, 2.70)</td>
<td>0.95</td>
<td>(0.36, 2.51)</td>
</tr>
<tr>
<td><strong>Phase CR3</strong></td>
<td>1.25</td>
<td>(0.24, 6.53)</td>
<td>1.38</td>
<td>(0.28, 6.76)</td>
</tr>
<tr>
<td><strong>Phase Relapse</strong></td>
<td>4.71</td>
<td>(2.11, 10.54)</td>
<td>4.06</td>
<td>(1.85, 8.92)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>1.89</td>
<td>(0.40, 8.99)</td>
<td>1.49</td>
<td>(0.32, 6.85)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>0.99</td>
<td>(0.97, 1.02)</td>
<td>0.99</td>
<td>(0.97, 1.02)</td>
</tr>
</tbody>
</table>
Discussion
We proposed a simple and flexible way of directly modeling the hazard function, using multinomial regression. This leads to smooth estimates of the absolute risks. We are explicitly modeling time. We can test the significance of covariates.
• We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
• We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
  • This leads to smooth estimates of the absolute risks.
Discussion

- We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
  - This leads to smooth estimates of the absolute risks.
- We are explicitly modeling time.
Discussion

• We proposed a simple and flexible way of directly modeling the hazard function, using multinomial regression.
  • This leads to smooth estimates of the absolute risks.
• We are explicitly modeling time.
• We can test the significance of covariates.
J. A. Hanley and O. S. Miettinen.  
**Fitting smooth-in-time prognostic risk functions via logistic regression.**  

O. Saarela.  
**A case-base sampling method for estimating recurrent event intensities.**  

O. Saarela and J. A. Hanley.  
**Case-base methods for studying vaccination safety.**  
L. Scrucca, A. Santucci, and F. Aversa.

Regression modeling of competing risk using R: an in depth guide for clinicians.

Questions or comments?

For more details, visit
http://sahirbhatnagar.com/casebase/