Bayesian Semi-Parametric Analysis of Semi-Competing Risks Data

Kyu Ha Lee, PhD
Sebastien Haneuse, PhD
Deborah Schrag, MD
Francesca Dominici, PhD

Harvard T.H. Chan School of Public Health
Harvard Medical School
Dana Farber Cancer Institute

July 11, 2019

Pancreatic Cancer

- Approximately 56,000 individuals will be diagnosed with pancreatic cancer in the U.S. this year
- Unfortunately there are no effective screening modalities
  - patients are usually diagnosed at late stages
- Large majority are not eligible for surgical treatment
  - chemotherapy is administered in the context of palliative care
- Prognosis is very poor
  - 5-year survival rate is 9%

Ongoing collaboration

- Broad goal is to characterize and understand variation in the quality of end-of-life care for patients diagnosed with pancreatic cancer
- Quality can be measured in many ways
- Our immediate focus is on readmission
  * readmission after discharge from the hospitalization at which the diagnosis was given
- Hospital-specific readmission rates are calculated and reported by CMS
  * logistic regression
  * Readmissions Reduction Program
  * Hospital Inpatient Quality Reporting Program
  * determine, in part, a hospitals reimbursement rate for the subsequent year

Challenges

Death as a competing risk

- Consider outcomes among $N = 16,051$ Medicare patients:
  - between 2005-2008
  - inpatient care claims, including hospitalizations

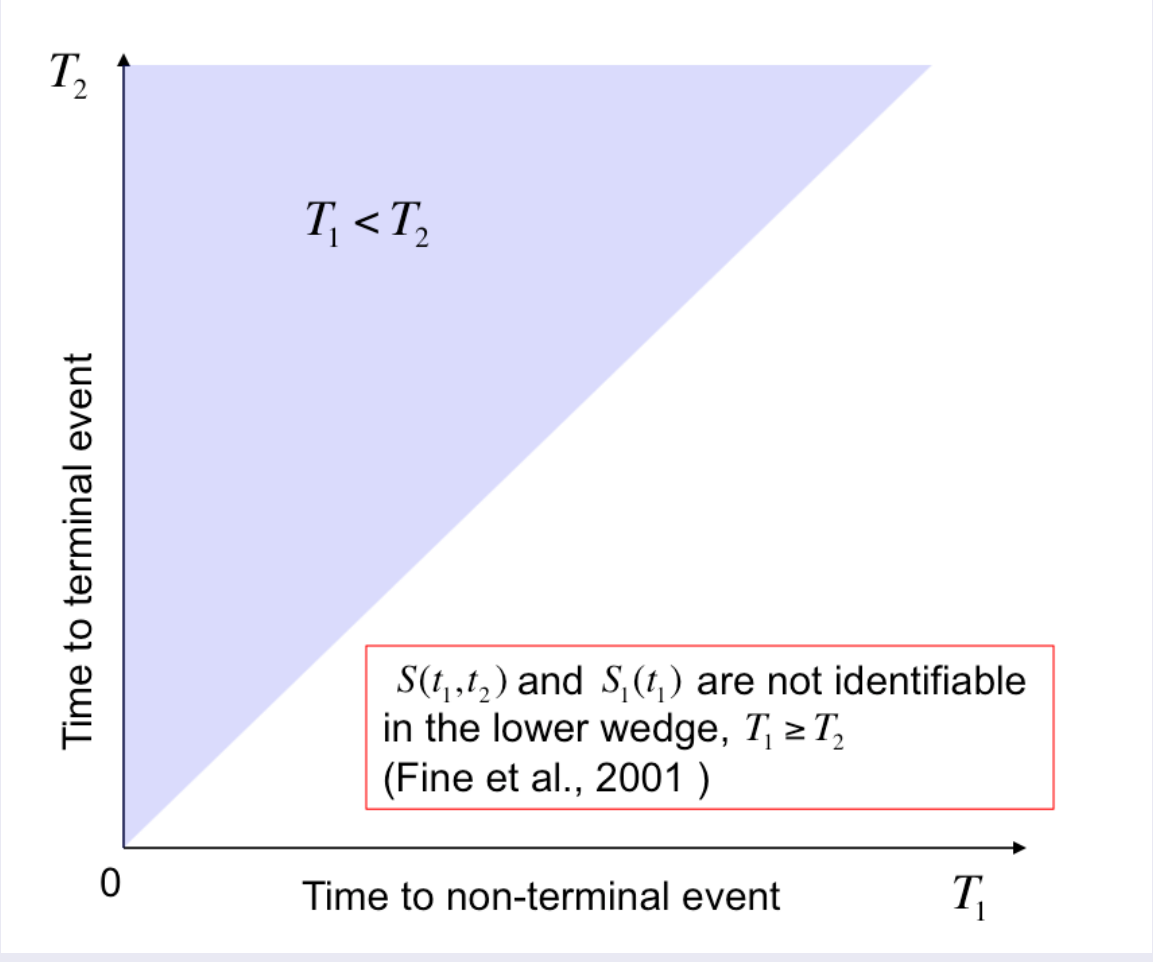
Observed events during the first 90 days		
Readmitted and subsequently died	2,254	14.0%
Readmitted and censored prior to death	2,213	13.8%
Death without readmission	7,505	46.8%
Censored prior to readmission or death	4,079	25.4%

- Primary interest lies with readmission or time-to-readmission
- We only observe readmission among folks who have not died
- Data that exhibit this structure is often referred to as *semi-competing risks* data

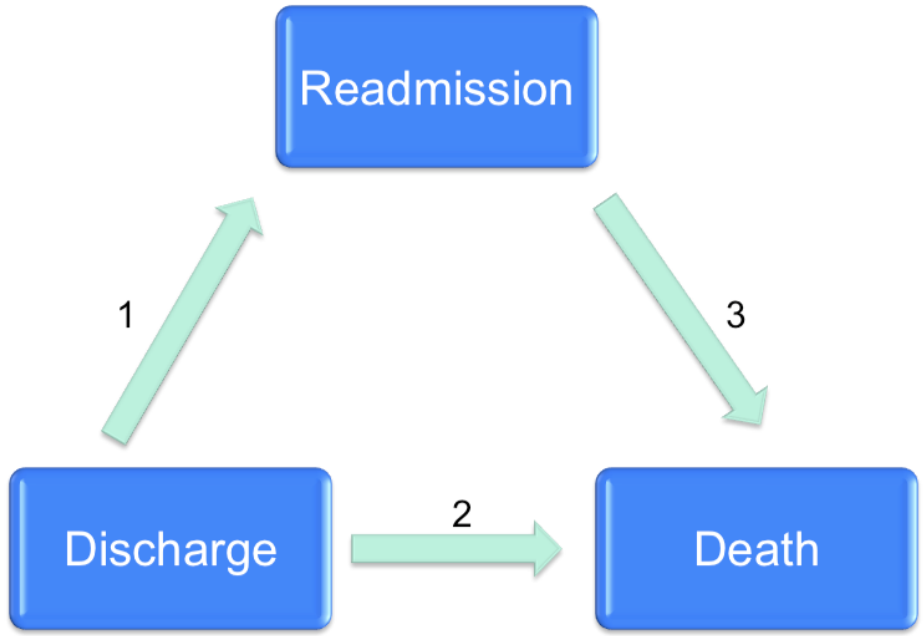
- Naïve approaches to learning about readmission might include:
  (a) logistic regression analyses
    * binary outcome
    * ignores death as a competing risk (!)
  (b) Cox regression analyses
    * time-to-readmission
    * treat death as an independent censoring mechanism (!)
  (c) composite endpoint analyses
    * first of either readmission or death
    * conflation changes the scientific question (!)
- In the profiling context, inappropriate handling of death may be problematic because:
  * a hospital may have a low readmission rates because they do a poor job of keeping patients alive
  * a hospital may have a high readmission rates because they do a good job of keeping patients alive

Semi-Competing Risks Problem

- If a patient dies prior to readmission, the time to readmission will never be observed.
- Key challenge: non-identifiability

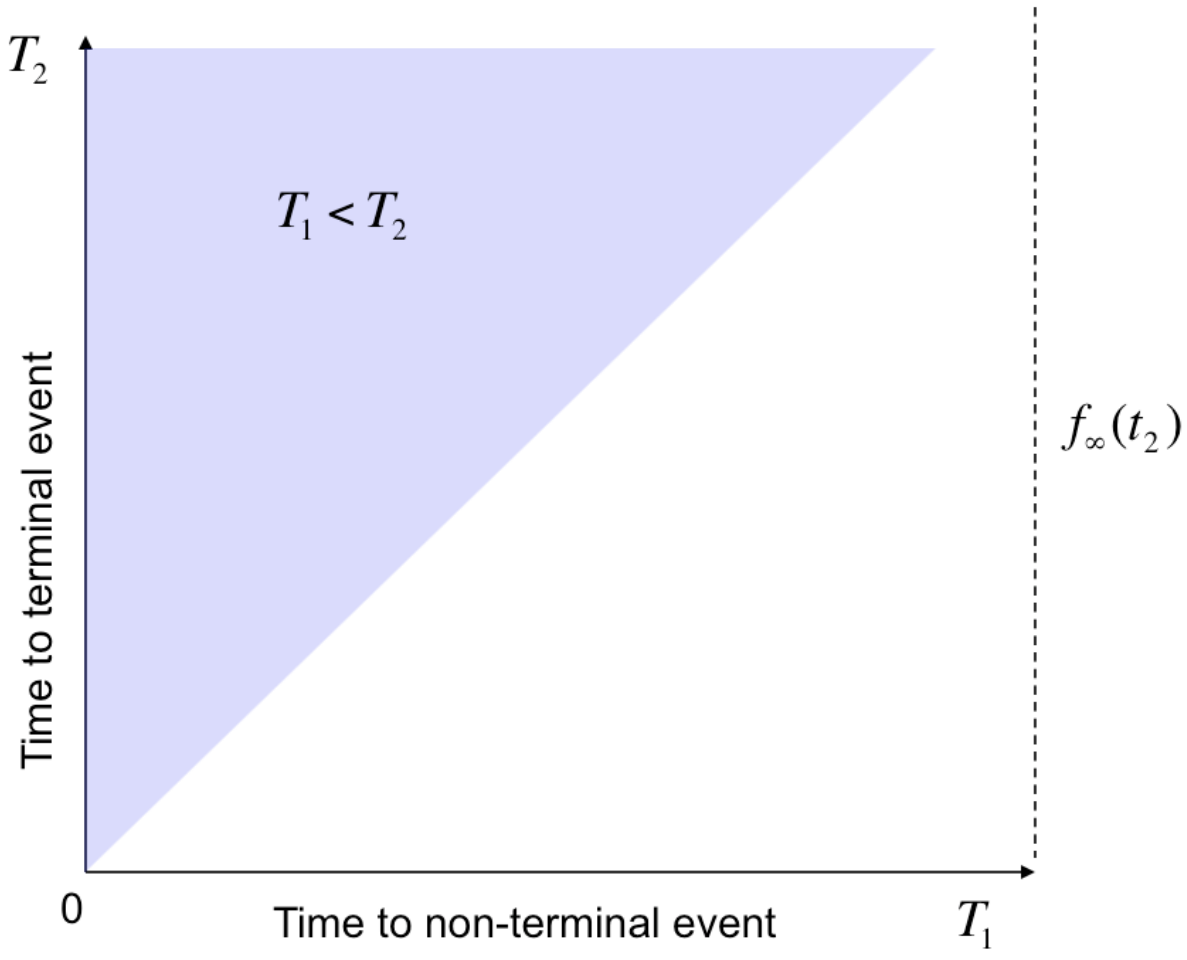


- An intuitive approach to analyzing semi-competing risks data is to view the data as arising from an underlying illness-death multi-state model.



- Movement between the states is governed by a set of transition-specific intensity or hazard functions

To permit the identifiability of the marginal density of T_1, Xu et al. (2010) set $T_1 = \infty$ if a subject experiences death prior to readmission.



Transition-Specific Hazard Functions

Modeling strategy is to place structure on the three hazard functions as follows:

$$
\begin{aligned}
h_1(t_{1i}|\gamma_i, \boldsymbol{x}_i) &= \gamma_i h_{01}(t_{1i}) e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_1}, \quad t_{1i} > 0, \\
h_2(t_{2i}|\gamma_i, \boldsymbol{x}_i) &= \gamma_i h_{02}(t_{2i}) e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2}, \quad t_{2i} > 0, \\
h_3(t_{2i}|t_{1i}, \gamma_i, \boldsymbol{x}_i) &= \gamma_i h_{03}(t_{2i}) e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}, \quad 0 < t_{1i} < t_{2i},
\end{aligned}
$$

* $\gamma_i \sim \text{Gamma}(\theta^{-1}, \theta^{-1})$ is a shared patient-specific frailty

- Our experience with both of these is that they are highly unstable
  - maximization (or root solving) over a large parameter space is tricky
- Potential benefits of the Bayesian paradigm:
  - ability to incorporate substantive prior information
  - automated quantification of uncertainty
  - prediction is straightforward
  - prescriptive nature of computation
- Four main challenges:
  (1) specification of the three continuous baseline hazard functions
  (2) prior elicitation and specification
  (3) robust and efficient computational schemes
  (4) user-friendly software

Baseline hazard functions

$$
\begin{aligned}
h_1(t_1) &= \gamma_{ji} h_{01}(t_1) \exp\left\{ \boldsymbol{X}_{ji1}^\top \boldsymbol{\beta}_1 \right\}, \quad t_1 > 0, \\
h_2(t_2) &= \gamma_{ji} h_{02}(t_2) \exp\left\{ \boldsymbol{X}_{ji2}^\top \boldsymbol{\beta}_2 \right\}, \quad t_2 > 0, \\
h_3(t_2|t_1) &= \gamma_{ji} h_{03}(t_2|t_1) \exp\left\{ \boldsymbol{X}_{ji3}^\top \boldsymbol{\beta}_3 \right\}, \quad 0 < t_1 < t_2.
\end{aligned}
$$

- One simple way forward would be to take the baseline hazard function from some parametric distribution
  - exponential/Weibull distribution
- Parametric modeling is often viewed in a negative light but it does have some advantages
  - estimation/inference tends to be (more) straightforward
  - typically more stable in data poor settings
  - prediction is more straightforward

- Nevertheless, towards a more flexible model specification, we also consider modeling each the logarithm of $h_{0g}()$ as a mixture of piecewise constant functions

$$\log(h_0(t)) = \lambda(t) = \sum_{j=1}^{J} 1_{[s_j < t \leq s_{j+1}]} \lambda_j$$

  * $\boldsymbol{s} = \{s_1, \ldots, s_J, s_{J+1}\}$ is a partition of the observed time scale

- Within the Bayesian framework we can treat J and $\boldsymbol{s}$ as 'random'
  * assign priors and update their values in the MCMC scheme
- Result is that the value of $\lambda(t)$ in any given small interval is (marginally) a mixture of piecewise constant functions
  * smooth!

The Observed Likelihood of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \boldsymbol{\gamma})$

Then the observed data likelihood for grouped or discretized survival times for n subjects has the following form in terms of the disjoint intervals:

$$\prod_{j=1}^{J_1+1} \prod_{k=1}^{J_2+1} \prod_{l=1}^{J_3+1} \exp\left\{\lambda_{1j} d_{1j} - e^{\lambda_{1j}} \sum_{m \in \mathcal{R}_{1j}} \Delta^1_{mj} \gamma_m e^{\boldsymbol{x}_m^\top \boldsymbol{\beta}_1}\right\}$$
$$\times \exp\left\{\lambda_{2k} d_{2k} - e^{\lambda_{2k}} \sum_{q \in \mathcal{R}_{1k}} \Delta^2_{qk} \gamma_q e^{\boldsymbol{x}_q^\top \boldsymbol{\beta}_2}\right\}$$
$$\times \exp\left\{\lambda_{3l} d_{3l} - e^{\lambda_{3l}} \sum_{r \in \mathcal{R}_{2l}} \Delta^{*3}_{rl} \gamma_r e^{\boldsymbol{x}_r^\top \boldsymbol{\beta}_3}\right\}$$
$$\times \prod_{m' \in \mathcal{D}_{1j}} \gamma_{m'} e^{\boldsymbol{x}_{m'}^\top \boldsymbol{\beta}_1} \prod_{q' \in \mathcal{D}_{2k}} \gamma_{q'} e^{\boldsymbol{x}_{q'}^\top \boldsymbol{\beta}_2} \prod_{r' \in \mathcal{D}_{3l}} \gamma_{r'} e^{\boldsymbol{x}_{r'}^\top \boldsymbol{\beta}_3},$$

The Prior Distributions

Our prior choices are, for $g \in \{1, 2, 3\}$:

$$
\begin{aligned}
\pi(\boldsymbol{\beta}_g) &\propto 1, \\
\boldsymbol{\lambda}_g | J_g, \mu_{\lambda_g}, \sigma^2_{\lambda_g} &\sim \mathcal{N}_{J_g+1}(\mu_{\lambda_g} \mathbf{1}, \sigma^2_{\lambda_g} \Sigma_{\lambda_g}), \\
J_g &\sim \mathcal{P}(\alpha_g), \\
\pi(\boldsymbol{s}_g | J_g) &\propto \frac{(2J_g + 1)! \prod_{j=1}^{J_g+1} (s_j - s_{j-1})}{(s_{g,J_g+1})^{(2J_g+1)}}, \\
\pi(\mu_{\lambda_g}) &\propto 1, \\
\sigma^{-2}_{\lambda_g} &\sim \mathcal{G}(a_g, b_g),
\end{aligned}
$$

and

$$
\begin{aligned}
\gamma_i | \theta &\sim \mathcal{G}(\theta^{-1},\ \theta^{-1}), \qquad i = 1, \ldots, n \\
\theta^{-1} &\sim \mathcal{G}(\psi,\ \omega).
\end{aligned}
$$

Computation and software

- MCMC via a random scan Gibbs sampling algorithm
- Most of the moves are straightforward
  * exploit conjugacies
  * Metropolis-Hastings update
- Certain moves for the baseline hazard functions requires a change in the dimension of the parameter space
  * those pertaining to the number of intervals, J
  * use a reversible-jump Metropolis-Hastings-Green update
- Implemented in the `SemiCompRisks` package for `R`
  * `C` is used as the primary work engine
  * documentation includes a series of cheat sheets specific to various models that might be of interest

Computational Scheme

- To perform posterior estimation and inference, we use a random scan Gibbs sampling algorithm to generate samples from the full posterior distributions.
- Updating $\boldsymbol{s}_g$ and J_g requires a change in the dimension of the parameter space; a reversible jump MCMC Metropolis-Hastings-Green (MHG) algorithm was developed and implemented (Green, 1995).
- R-package "SemiCompRisks" is available in CRAN.

3. Application

- The data available for this study consists of information on 100% Medicare enrollees from Jan/2005 to Nov/2008.
- A total of 16,051 individuals aged 75 years or older are considered:
  1. They were hospitalized with a diagnosis of pancreatic cancer,
  2. They did not undergo any pancreatic cancer specific procedures (i.e. their disease was sufficiently advanced that curative treatment was not a viable option).
- For both outcomes, we (administratively) censored observation time at t = 90 days.

Objectives

- Identifying risks factors for time to readmission
- Providing the measure of dependence between time to readmission and time to death
- Predictive probability of being readmitted

Table: Posterior medians (PM) and 95% credible intervals (CI) for hazard ratio parameters ($\exp(\boldsymbol{\beta}_g)$, $g \in \{1, 2, 3\}$) from semi-competing risks analyses based on the proposed Bayesian framework. Results are based on setting the Poisson rate parameters α_g, $g \in \{1, 2, 3\}$, to 20 for all MVN-ICAR specifications of baseline hazard functions.

		Readmission	Death prior to readmission	Death after readmission
		PM (95% CI)	PM (95% CI)	PM (95% CI)
Comorbidity index[a]	0-1	1.00	1.00	1.00
	2-3	1.03 (0.96, 1.12)	0.99 (0.93, 1.05)	0.99 (0.89, 1.10)
	≥ 4	1.26 (1.16, 1.37)	1.15 (1.07, 1.23)	1.07 (0.95, 1.21)
Race	White	1.00	1.00	1.00
	Non-white	1.27 (1.17, 1.39)	0.86 (0.79, 0.93)	1.13 (1.01, 1.28)
Gender	Female	1.00	1.00	1.00
	Male	1.10 (1.03, 1.18)	1.30 (1.23, 1.38)	1.22 (1.12, 1.34)
Age[b]		0.87 (0.84, 0.90)	1.07 (1.04, 1.10)	1.08 (1.03, 1.13)
Care after discharge	Home	1.00	1.00	1.00
	Home care	1.21 (1.12, 1.31)	1.53 (1.39, 1.69)	1.23 (1.10, 1.38)
	ICF/SNF	0.82 (0.75, 0.91)	3.46 (3.19, 3.79)	1.76 (1.54, 2.01)
	Hospice	0.18 (0.15, 0.21)	8.96 (8.25, 9.86)	3.08 (2.38, 3.99)
Hospital stay	≤ 2 weeks	1.00	1.00	1.00
	> 2 weeks	1.25 (1.12, 1.39)	1.09 (1.00, 1.20)	0.89 (0.76, 1.05)

[a] Number of diagnosis codes given during the initial hospitalization from a list of 27 disease/disorders related to prognosis following hospital discharge.

[b] Standardized so that a one-unit contrast corresponds to a difference of 5 years.

Identifying Risk Factors for Time to Readmission

- There is evidence of increased risk for readmission associated with a high comorbidity index, a long (initial) hospital stay, non-white race, male gender, and discharge to home care.

- The semi-competing risks analysis reveals nuance in how several covariates confer risk for death.
  - The semi-competing risks analysis reveals the evidence of the decreased risk of death for an individual with non-white race who have not been readmitted (HR 0.86; 95% CI 0.79, 0.93) and the evidence of the increased risk of death for an individual with non-white race after readmission (HR 1.13; 95% 1.01, 1.28).

Table: Covariate profiles of the four different individuals considered for the explanatory hazard ratio and the posterior predictive distribution

	Comorbidity index	Race	Gender	Age	Care after discharge	Hospital stay
Baseline	0-1	White	Female	82	Home	≤ 2 weeks
Subject 1	≥ 4	Non-white	Male	92	Home care	> 2 weeks
Subject 2	0-1	Non-white	Female	92	Home	≤ 2 weeks
Subject 3	≥ 4	White	Male	82	Hospice	> 2 weeks

Measure of dependence between T_1 and T_2

Explanatory Hazard Ratio (EHR):

$$\frac{h_3(t_2|t_1,\gamma,\boldsymbol{x})}{h_2(t_2|\gamma,\boldsymbol{x})} = \frac{h_{03}(t_2)}{h_{02}(t_2)} \exp[\boldsymbol{x}^\top(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2)]$$

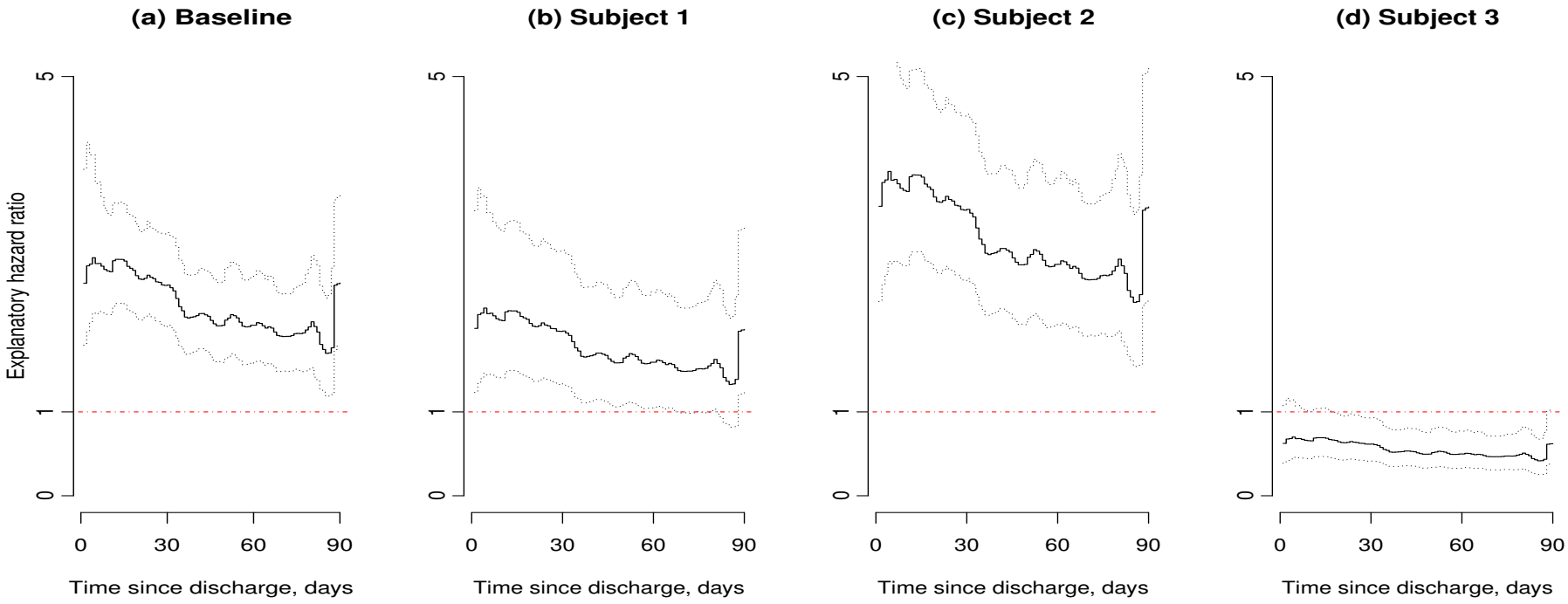


Figure: Pointwise posterior median and 95% credible intervals for the explanatory hazard ratio (EHR) for the four individuals.

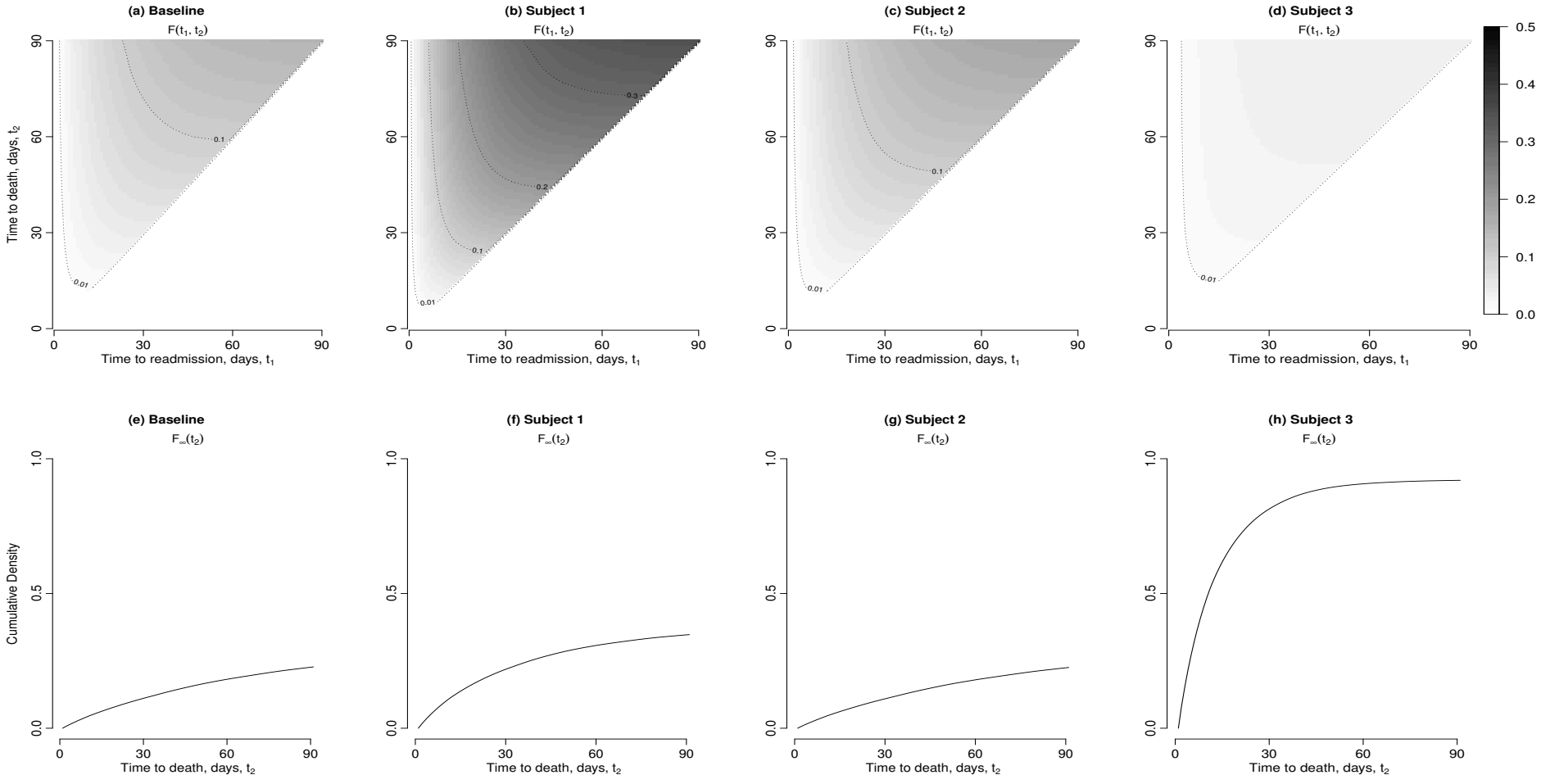


Figure: Posterior predictive distribution (conditioning on $\gamma = 1$) for four individuals; panels (a)-(d) show the posterior predictive distribution $F(t_1, t_2)$ for $t_1 \leq t_2$; panels (e)-(h) provide the posterior predictive distribution $F_\infty(t_2)$.

Final comments

- Semi-competing risks framework provides an opportunity to think about any given line of research in a different way.
  - consider the two events jointly
- Implemented in the `SemiCompRisks` package for `R`
- Cluster-correlated data (JASA, 2016), AFT model (Biometrics, 2017)

Acknowledgements

- Sebastien Haneuse
- Francesca Dominici
- Deborah Schrag
- Yun Wang

- This work was supported by National Institutes of Health grants (R01 CA181360-01, P01 CA134294-02, ES012044, K18 HS021991).

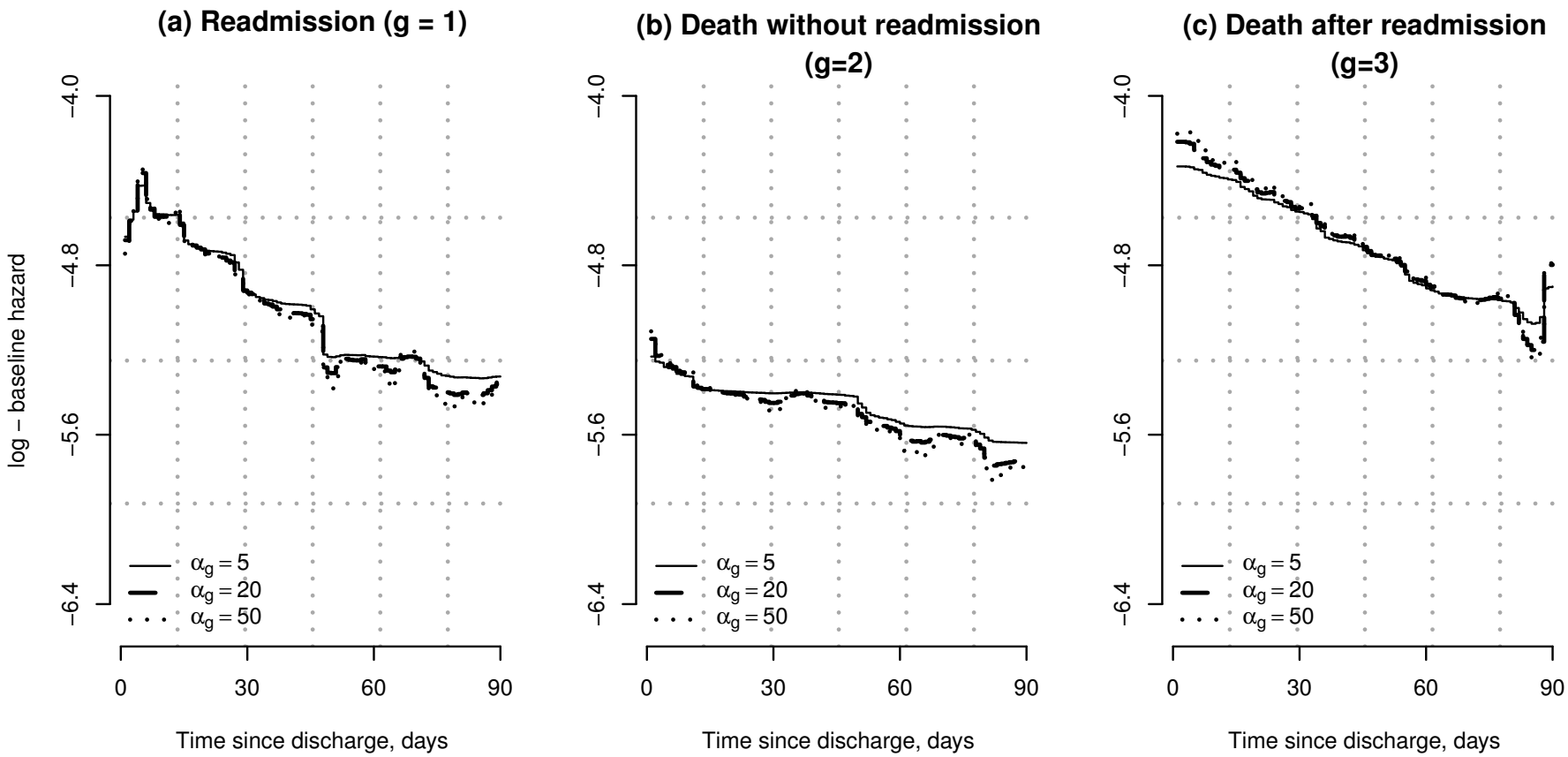


Figure: Estimates of the log-baseline hazard functions from the proposed Bayesian framework for semi-competing risks analysis. Three sets of analyses were performed, with values of α_g of 5, 20 and 50 adopted for all Poisson rate parameters.