Semi-competing risks: Accounting for death as a competing risk in public health research when the outcome of interest is non-terminal

Sebastien Haneuse, PhD
Harrison Reeder, AM

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Overview of the workshop

Who were are

- Sebastien Haneuse
- Professor of Biostatistics

- Harrison Reeder
- 4th-year PhD student

Structure

- Outline:
  - preeclampsia
  - semi-competing risks
  - data analysis methods
  - software
  - worked example
  - additional topics
  - references

- Breakdown of the 2 hours
  - 1h15m of methods
    - 5-minute break sometime close to the top of the hour
  - 30m for the worked example application
  - 15m to finish off & field questions

- Adjustments:
  - online and down from 3 hours to 2
  - no hands-on software demonstration
  - code for the analyses we present is available

- What is expected of you?
  - nothing is specifically expected but some things will be helpful
  - understanding of core concepts in time-to-event data analysis
    - mainly the notion of (right) censoring
  - familiarity with concepts related to longitudinal or cluster-correlated data analysis
    - notion of 'dependence'
    - mixed effects models

Preeclampsia

- Preeclampsia is a condition that arises during pregnancy, one that is characterized by
  - high blood pressure
  - swelling in the hands and feet
  - excess protein in the urine
  - a range of other signs and symptoms
- Diagnosed following the 20-week gestation mark
- Affects 2-8% of pregnancies globally
- Leading cause of maternal and perinatal mortality
  - 16% of maternal deaths worldwide attributed to preeclampsia and related hypertensive disorders

- Can lead to a range of complications, including:
  - fetal growth restriction
  - premature delivery
    - and resulting complications
  - maternal eclampsia
    - seizures that occur during a woman's pregnancy or shortly after giving birth
  - HELLP syndrome

- Substantial costs
  - Stevens et al (*AJOG*, 2017) estimated cost of PE within the first 12 months of delivery in 2012 was ∼$2.2 billion

Risk factors

- There are no known causes of preeclampsia
- Numerous factors have been associated with risk, with varying degrees of strength of evidence
  - Giannakou et al (*UOG*, 2018)
  - Townsend et al (*UOG*, 2019)
- Maternal demographics
  - age
  - race/ethnicity
- Characteristics of the pregnancy
  - primiparity
  - in-vitro fertilization

- Maternal clinical characteristics
  - history of preeclampsia, hypertension and/or familial preeclampsia
  - obesity, diabetes and chronic kidney disease
  - smoking
  - polycystic ovary syndrome
  - mental stress
  - pregnancy-associated plasma protein-A
  - serum iron levels

Management

- The only cure for preeclampsia is to give birth
  - even after delivery, symptoms can last 1 to 6 weeks
- During pregnancy, preeclampsia can often be managed until the baby is sufficiently mature to be delivered
  - medications to lower blood pressure
  - anti-convulsive medications
  - bed rest
  - careful monitoring
- Requires balancing the risks of
  - early delivery for the baby
  - continued preeclampsia symptoms for the mother

A hypothetical study

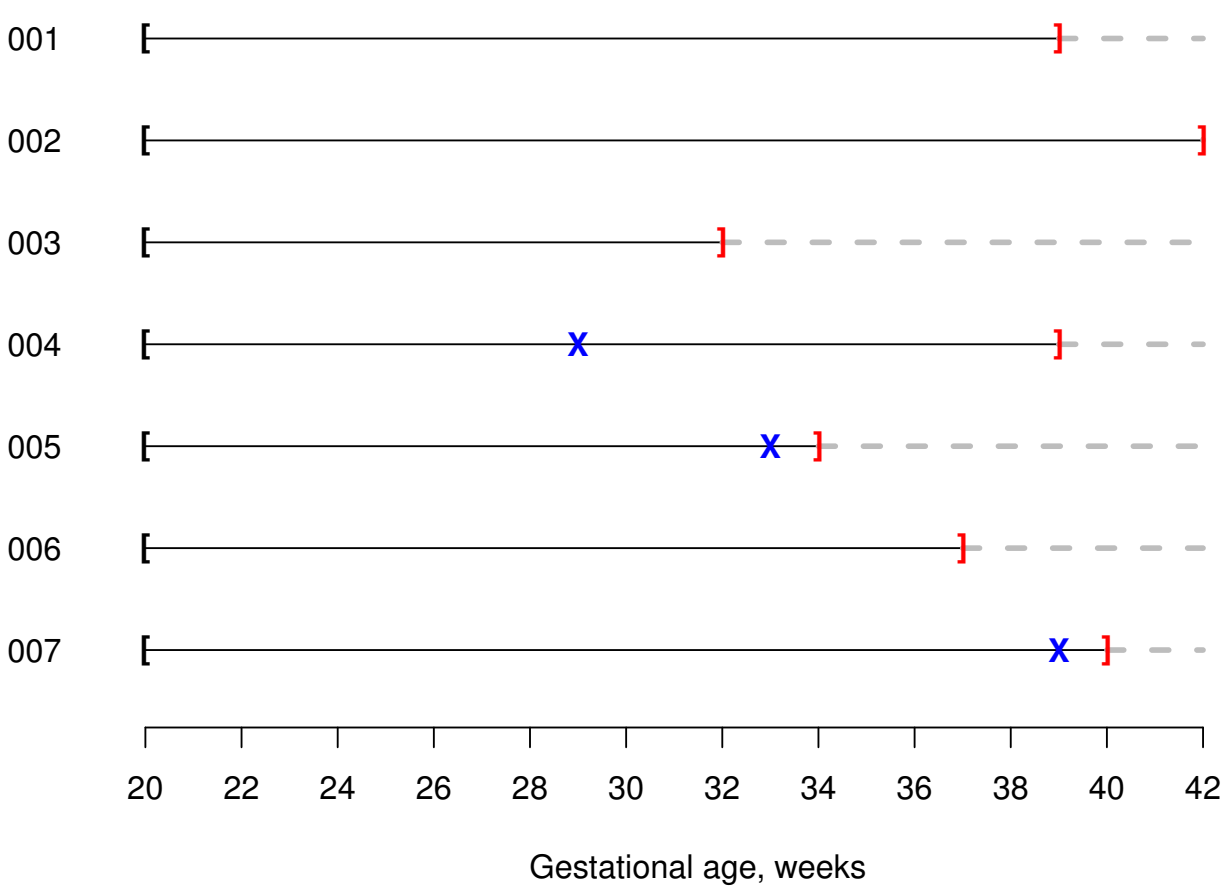
- Suppose we are interested in conducting a hypothetical study of preeclampsia, perhaps to
  - assess the effect of a novel treatment, or
  - investigate the association between some novel biomarker and risk, or
  - develop a novel risk prediction tool
- Consider the outcome to be '*incident clinical diagnosis of preeclampsia*'
- Operationally, we might specify the outcome to be binary indicator
  - $Y = 0/1 =$ no diagnosis/diagnosis
  - then use, say, logistic regression as a modeling framework
- Before forging ahead, it is worth looking at the data in some detail

- Some example (i.e. made-up) data:

ID	Age	HTN	BMI	Smoker	PE	Delivery	Y
001	28	N	36	former	NA	39	0
002	32	N	28	never	NA	42	0
003	23	N	27	never	NA	32	0
004	37	Y	41	former	29	39	1
005	34	Y	34	current	33	34	1
006	19	N	25	former	NA	37	0
007	41	N	29	never	39	40	1
...							

  * preeclampsia (PE) and delivery are measured in weeks of gestation
  * PE = NA indicates no that there was no diagnosis
  * no right censoring here

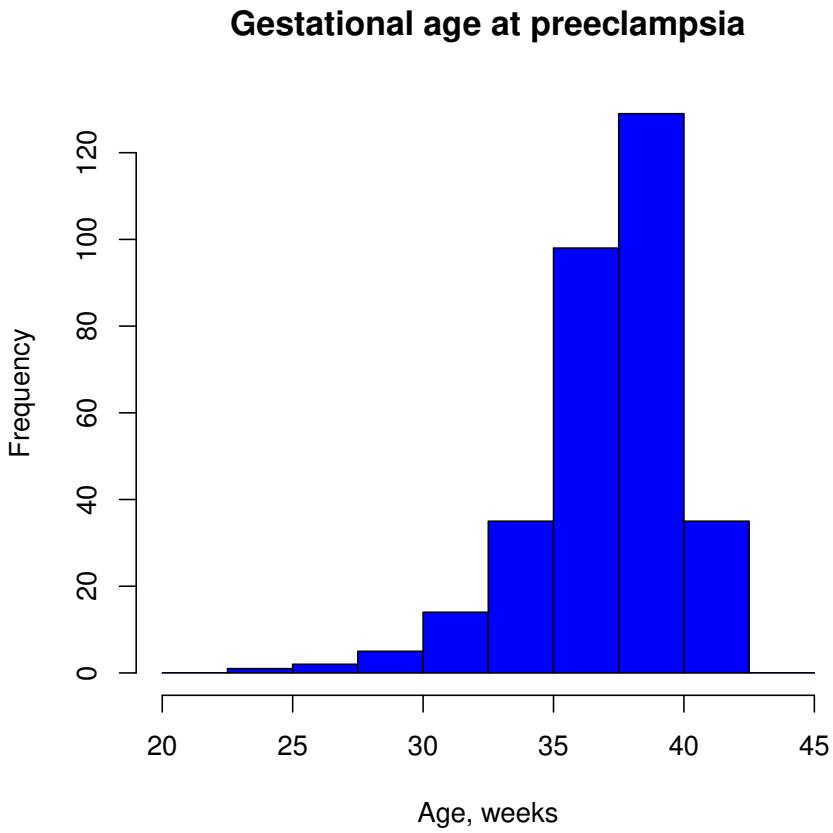
- Graphically, the outcome information can be represented as:



Q: What do we make of the heterogeneity in:

* the amount of person-time at-risk for preeclampsia?
* the timing of preeclampsia among those with a diagnosis?

- Both types of heterogeneity exist across N=5,054 women from BIDMC:
  - all singleton births in 2016
  - n=319 (6.3%) had a diagnosis of preeclampsia



  - more on these data later

- Suppose we ignore all of this heterogeneity and forge ahead with a logistic regression analysis of the binary outcome, Y

- This analysis would ignore the timing of the diagnosis of preeclampsia
  - ignore potentially useful information
  - results in a limiting of the scope of enquiry
  - cannot, for example, investigate whether a treatment works to delay a diagnosis

- Such an analysis would also ignore the fact that the amount of person-time during which a pregnant mother is at-risk to be diagnosed with preeclampsia varys
  - combine 0/1 outcomes defined over a 20-35 week window with outcomes defined over a 20-42 week window

Q: Conceptually, does it make sense to do this? Are we not combining 'apples' and 'oranges' (and 'plums' and ...)?

- More pernicious, however, is that the analysis would (erroneously) assume that the amount of person-time at-risk is the same across all women in the dataset
  - a logistic regression simply doesn't know any different
- Unclear how to interpret the results
  - mixing of effects

Time-to-event analysis

Q: Could we, instead, make progress with a time-to-event (or survival) analysis?

- outcome is '*timing of incident clinical diagnosis of preeclampsia*'
- treat the outcome as being *censored* if a woman delivers first

- Observed outcome data become:

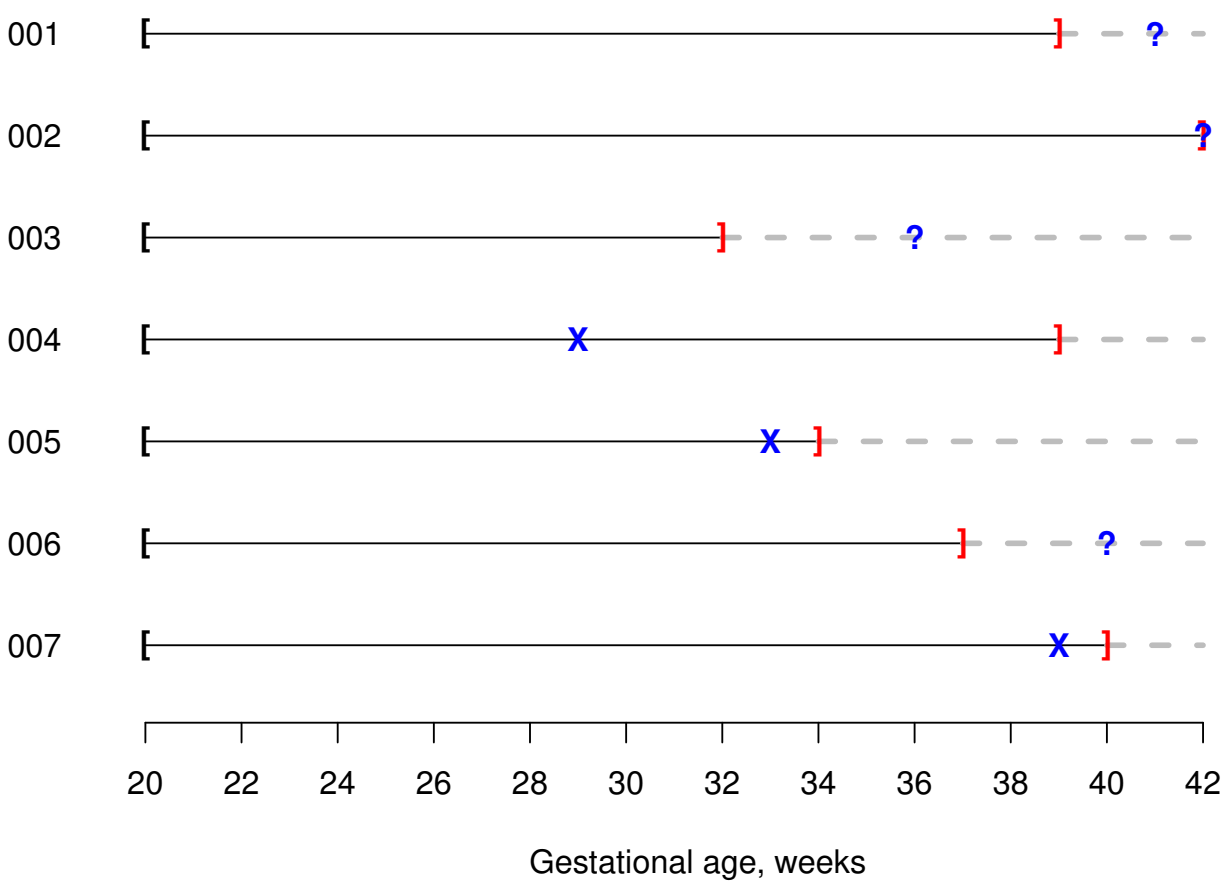
ID	Age	HTN	BMI	Smoker	PE	Delivery	Y	delta
001	28	N	36	former	NA	39	39	0
002	32	N	28	never	NA	42	42	0
003	23	N	27	never	NA	32	32	0
004	37	Y	41	former	29	39	29	1
005	34	Y	34	current	33	34	33	1
006	19	N	25	former	NA	37	37	0
007	41	N	29	never	39	40	39	1
...								

- Could then make use of the broad range of well-known statistical methods, such as:
  - Kaplan-Meier plots of the survivor function
  - log-rank hypothesis testing
  - Cox models for the hazard function

- Problematic with going down this path, however, is that censoring in time-to-event analyses is a phenomenon that is specific to the capacity of the research team to *observe* an event

- Moreover, standard time-to-event analyses implicitly assume:
  - individuals who are censored remain at-risk to subsequently experience the event
  - that is, individuals will eventually experience the event

- Put another way, the issue with censoring is that we just don't get to see *when* they experience the event



  - essentially a missing data problem

Q: Does it make sense to view 'delivery' through this lens?

- preeclampsia is a condition that is specific to pregnancy
- following delivery a woman might be diagnosed with incident hypertension but not preeclampsia

- It does not seem reasonable to conceive of a woman as continuing to be at risk for preeclampsia following delivery
  - it is not even clear whether it makes sense to think of 'risk' of preeclampsia

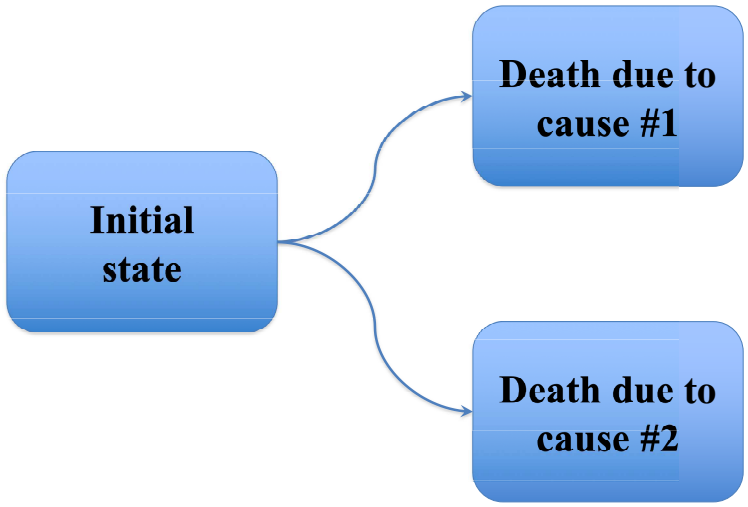
- Even if we were willing to ignore this conceptual issue, it is important to remember that statistical methods for time-to-event data assume *independent censoring*
  - intuitively, the assumption says that the experience with respect to the event of interest among those individuals who are censored would have been the same as those who were not been censored

- In the present context, the assumption essentially states that the timing of delivery provides no information about the timing of preeclampsia
  - it seems intuitive, however, that the timing of delivery and the timing of preeclampsia are not independent

Competing risks analysis

- Arguably, a more appropriate framing for the impact that 'delivery' exerts is as a *competing risk*
- Typically applied in the content of mortality
  - e.g. interest lies in cancer-specific mortality but one has to acknowledge other causes of death
- Apply a range of methods that are (increasingly) well-known, including:
  - cause-specific cumulative incidence functions
  - cause-specific hazard regression models
  - models for the sub-distribution hazard function
    - the 'Fine and Gray method'

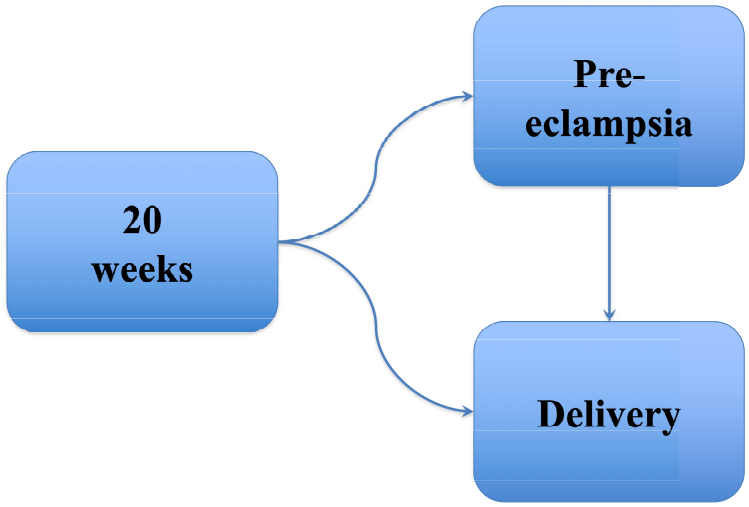
- In the usual set-up, each competing risk competes with each of the others



  * no subsequent transitions, regardless of which event arises first

- In the context of preeclampsia, however, while we might reasonably conceive of delivery as a competing risk, the reverse is not the case
  * a woman continues to be 'at risk' for delivery even if she has been diagnosed with preeclampsia

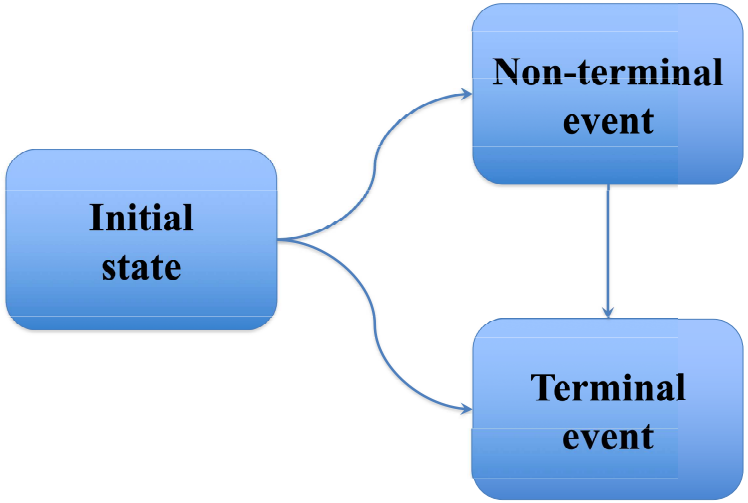
- Results in an asymmetry between the two events of preeclampsia and delivery



- This asymmetry is sometimes referred to as *semi-competing risks*
  - Fine et al (*Biometrika*, 2001)

Semi-competing risks

- Setting where interest lies in a *non-terminal event* that is subject to a *terminal event*
  - begin in some 'initial' state
  - terminal event is a competing risk for the non-terminal event
  - but <u>not</u> vice versa



- In the running context:
  - the initial state is being pregnant at 20 weeks
  - preeclampsia is non-terminal
  - delivery is terminal

- Beyond preeclampsia, semi-competing risks arise in a wide range of settings
- Some specific areas that our work has taken us to:

Clinical domain	**Non-terminal event(s)**	**Terminal event**
Pregnancy	Preeclampsia	Delivery
The elderly	Alzheimer's & dementia	Death
Palliative care at the end of life	Readmission	Death
HSCT recipients	GVHD & relapse	Death
ICD following heart failure	Shock	Death
Preterm infants	NEC, IVH & discharge	Death

- Emphasize that the framing is really only useful when the non-terminal event is of primary interest as an outcome
  - either individually or jointly with the terminal event

- If primary interest lies in the terminal event as the outcome, with the non-terminal event possibly a time-varying covariate, then one can (and should) use
  - time-to-event analysis methods
  - methods for competing risks

Methodologic considerations

- Summarizing the discussion so far, the use of standard (well-known) statistical methods when analyzing semi-competing risks data generally fail in one or both of the following:

  (1) respecting the competing risk role that terminal event plays

  (2) acknowledging the potential for dependence between non-terminal and terminal event

- Need statistical analysis methods that resolve these issues while also being able to handle the other usual suspects:
  - measurement error
  - missing data in covariates
  - various forms of censoring/truncation
    - right censoring
    - left truncation
    - interval censoring

Scientific considerations

- That we observe both the non-terminal event and the terminal event on at least some individuals means that we can learn about:
  - dependence between the two events
  - how the two events covary over time as a function of covariates
- Also provides an opportunity to develop joint prediction tools that simultaneously consider risk of the two events
  - clinically, it seems that any conversation about the risk and timing of preeclampsia would also involve discussion of the timing of delivery
- Neither univariate or competing risk analyses methods make use of this information
  - ignoring this information, as these methods do, represents a lost opportunity from a scientific perspective

The analysis of semi-competing risks data

- In its most general form, there are four possible outcome scenarios for patients during the observed person-time period:

  (1) experienced the non-terminal event and then censored

  (2) experienced both, that is the non-terminal event followed by the terminal event

  (3) experienced the terminal event, without having experienced the non-terminal event

  (4) censored prior to experiencing either event

- In regard to the covariate data, we only consider settings where 'baseline' covariates are of interest
  - that is, we do not consider time-varying covariates

- Notation for the outcome data:
  - time of the non-terminal event, T_1
  - time of the terminal event, T_2
  - censoring time, C

- Observed data for the **non-terminal event**:

$$Y_1 = \text{minimum of } T_1, T_2 \text{ and } C$$

$$\delta_1 = \begin{cases} 1 & \text{if observed to experience the non-terminal event} \\ 0 & \text{otherwise (i.e. terminal or censored first)} \end{cases}$$

- Observed data for the **terminal event**:

$$Y_2 = \text{minimum of } T_2 \text{ and } C$$

$$\delta_2 = \begin{cases} 1 & \text{if observed to experience the terminal event} \\ 0 & \text{otherwise (i.e. censored first)} \end{cases}$$

Major threads

- Three major threads for the statistical analysis of semi-competing risks data:

  (1) methods grounded in the causal inference paradigm

  (2) copula-based methods

  (3) the illness-death model

- Much of what is presented here will fall into thread (3)
  - brief overview of threads (1) and (2)

- Comprehensive set of references at the end of these notes

Methods grounded in the causal inference paradigm

- Suppose interest lies in comparing the relative impact of two (or more) treatment options on the risk of the non-terminal event

- Overarching strategy of the causal inference paradigm is to:
  - (i) define a causal contrast of interest
  - (ii) specify identifying conditions
  - (iii) develop an estimator and establish its properties

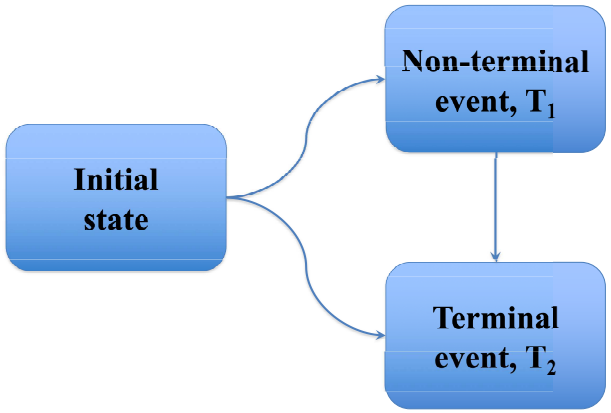
- One such contrast is the *survivor averaged causal effect* (SACE)
  - the impact of treatment choice among individuals in the 'always survivors' principal stratum
  - patients who would survive under either treatment
    - at least over some well-defined and meaningful window of follow-up
  - sub-population for whom the non-terminal event is always well-defined

- Key methodologic challenge is that this stratum membership is not known, so that SACE is not fully identified from the data
  - reliance on an untestable assumption
- Various methods have been developed for
  - constructing bounds
  - performing sensitivity analyses
- Conceptually, this way forward suffers in two ways:
  - careful consideration is needed in regard to what is meant by the 'always survivors' principle stratum

    Q: who are these individuals? can they be 'identified' in the general population?
  - precludes learning about dependence between the two events
    - opportunity lost

Copula-based methods

- Recall that T_1 and T_2 denote the times to the non-terminal and terminal events, respectively



- Copula-based methods for semi-competing risks build on corresponding methods for standard bivariate time-to-event analyses
  - where neither event is terminal for the other
  - e.g. joint survival among twins

- Specifically, they focus on learning about T_1 and T_2 via the *joint survivor function*: $P(T_1 > t_1, T_2 > t_2)$

- Strategy in standard bivariate settings:

(i) specify models for the marginal survivor functions:

$$S_1(t_1) = P(T_1 > t_1)$$

and

$$S_2(t_2) = P(T_2 > t_2)$$

(ii) 'link' these functions to form a model for the joint survivor function:

$$P(T_1 > t_1, T_2 > t_2) \;=\; C_\theta(S_1(t_1), S_2(t_2))$$

* $C_\theta(\cdot, \cdot)$ is referred to as a *copula*
* essentially a mathematical function
* θ is an unknown parameter

(iii) simultaneously estimate the components of $S_1(t_1)$, $S_2(t_2)$ and θ

- One well-known copula is the *Clayton copula*

$$C_\theta(S_1(t_1), S_2(t_2)) \;=\; \left\{S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right\}^{1/(1-\theta)}$$

  - motivated by consideration of a latent *subject-specific frailty*
    - assumed to impact the risk of both events
    - much as random effects do in mixed effects models
  - induces a form of dependence between the two outcomes
  - θ is the variance of the frailties
    - quantification of the magnitude of dependence that arises via this 'shared' mechanism

- One general limitation of applying this strategy to semi-competing risks, however, is that $S_1(t_1)$ cannot be interpreted as a marginal survival function
  - the marginal distribution of T_1 is not identified
  - unclear how to interpret the output from such an analysis

Illness-death models

- Illness-death models posit that, at any given point in time, the study participant is in one of the three 'states':

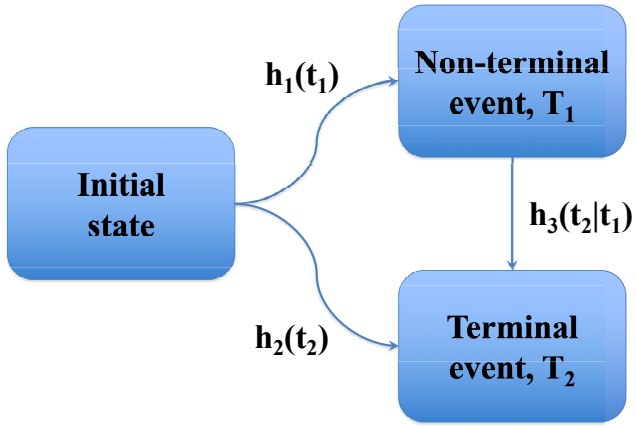
  (1) an initial state, prior to experiencing either event;

  (2) a state of having experienced the non-terminal event, necessarily prior to experiencing the terminal event; and,

  (3) an absorbing state of having experienced the terminal event

- Special case of a more general class of *multi-state models*

- Conceive of three *hazard functions* that dictate the rates at which individuals transition between the states



- Formally:
  - $h_1(t_1)$ is the cause-specific hazard for the non-terminal event, conditional on neither event having had occurred

$$h_1(t_1) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathsf{P}(T_1 \in [t_1, t_1 + \Delta) | \; T_1 \geq t_1, T_2 \geq t_1)$$

  - $h_2(t_2)$ is the cause-specific hazard for the terminal event, conditional on neither event having had occurred

$$h_2(t_2) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathsf{P}(T_2 \in [t_2, t_2 + \Delta) | \; T_1 \geq t_2, T_2 \geq t_2)$$

  - $h_3(t_2|t_1)$ is the cause-specific hazard for the terminal event, conditional on the non-terminal event having had occurred at time t_1

$$h_3(t_2|t_1) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathsf{P}(T_2 \in [t_2, t_2 + \Delta) | \; T_1 = t_1, T_2 \geq t_2)$$

- Towards answering a particular question of interest, we can then place model structure on each of these hazard functions

A general Cox-type illness-death model

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, && t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, && t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, && t_2 > t_1
\end{aligned}
$$

* transition-specific covariates
  * $\boldsymbol{X}_{1i}$, $\boldsymbol{X}_{2i}$ and $\boldsymbol{X}_{3i}$
* transition-specific covariate effects
  * $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$
* transition-specific baseline hazard functions:
  * $h_{01}(\cdot)$, $h_{02}(\cdot)$ and $h_{03}(\cdot)$
* subject-specific *shared frailty*
  * γ_i, for $i = 1, \ldots, n$

Cox-type model: transition-specific covariates

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, && t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, && t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, && t_2 > t_1
\end{aligned}
$$

- Permits different covariates to be taken to operate on the two outcomes
- For example, cervical length
  - has no known association with preeclampsia
  - is strongly associated with preterm delivery

Cox-type model: transition-specific covariate effects

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, & t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, & t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, & t_2 > t_1
\end{aligned}
$$

- Interpretations analogous to those in a standard Cox model
  - i.e. log hazard ratio
- Care to fold in the precise nature of the hazard
  - i.e. $h_1(t_1)$ is the *cause-specific hazard for the non-terminal event*
- The effect of a given covariate is permitted to be different across the three transitions
- For example, a novel prophylactic strategy may be hypothesized to reduce the risk of preeclampsia but possibly bring forward delivery

Cox-type model: transition-specific baseline hazard functions

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, && t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, && t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, && t_2 > t_1
\end{aligned}
$$

- Each has the usual interpretation of the hazard, as a function of time
  - i.e. among patients with $\boldsymbol{X}_{1i} \equiv \boldsymbol{0}$, $\boldsymbol{X}_{2i} \equiv \boldsymbol{0}$ or $\boldsymbol{X}_{3i} \equiv \boldsymbol{0}$ (as appropriate)
- Note, as currently specified, $h_{03}(t_2|t_1)$ is distinct for each t_1
  - has the potential to be quite complex
- Typically make progress by making some simplifying assumption about how $h_{03}(t_2|t_1)$ depends on t_1
  - more on this soon

Cox-type model: subject-specific shared frailties

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, && t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, && t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, && t_2 > t_1
\end{aligned}
$$

- Earlier we intuited that T_1 and T_2 may often reasonably be held to be dependent
- View T_1 and T_2 as subject-specific 'multivariate data'
- Build on familiar methods in the analysis of other types of multivariate data
  - i.e. longitudinal or clustered data
- Here, the γ_i have a role, interpretation and specification that are analogous to random effects in mixed effects models

- Shared factor that 'links' the specification of a subjects' transition-specific hazard functions
  - represents a summary of the collective impact of factors not included in the model
  - latent and not observed

- Serves to account for a specific type of (potential) dependence between T_1 and T_2

Cox-type model: estimation and inference

- Suppose we omit the subject-specific frailties from the model specification:

$$\begin{aligned} h_1(t_1) &= h_{01}(t_1)\ \exp\{\boldsymbol{X}_{1i}^T\boldsymbol{\beta}_1\}, & t_1 > 0 \\ h_2(t_2) &= h_{02}(t_2)\ \exp\{\boldsymbol{X}_{2i}^T\boldsymbol{\beta}_2\}, & t_2 > 0 \\ h_3(t_2|t_1) &= h_{03}(t_2|t_1)\ \exp\{\boldsymbol{X}_{3i}^T\boldsymbol{\beta}_3\}, & t_2 > t_1 \end{aligned}$$

- Simplification results in being able to estimate $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ as one would in a standard univariate setting
  - i.e. without the need to specify (or even consider) the baseline hazard functions
  - use standard software tools
    - `coxph` in `R`
    - `PROC PHREG` in `SAS`
    - `stcox` in `stata`

- Suppose, however, retaining the frailties is viewed as important/necessary, as in:

$$\begin{aligned} h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, & t_1 > 0 \\ h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, & t_2 > 0 \\ h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, & t_2 > t_1 \end{aligned}$$

  - sense/belief that they are needed to adequately characterize dependence
  - understanding dependence is of intrinsic interest

- Unfortunately, one cannot use standard methods to estimate $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$

- Moreover, it becomes necessary to provide concrete specifications for
  - the baseline hazard functions
  - the distribution of the subject-specific frailties

- For the frailties, it is typical to assume that they arise from a Gamma(θ^{-1}, θ^{-1}) distribution
  - E$[\gamma_i] = 1$ and Var$[\gamma_i] = \theta$
    - θ quantifies heterogeneity across individuals
  - choice has some important analytic benefits
    - important because with at-most 2 'observations' per subject there is not a huge amount of information about individual γ_i
- For the baseline hazard functions there is substantially more choice:
  - specifications based on a parametric distribution
    - Exponential, Weibull, Gompertz, etc
  - flexible spline-based specifications
    - P-splines, B-splines, Royston-Parmar splines etc
  - Bayesian non-parametric specifications
    - Dirichlet process mixtures

More on dependence

- Just as in the analysis of longitudinal data, that T_1 and T_2 are potentially dependent may be a statistical nuisance or of intrinsic interest

- Prior to modeling, it is important to acknowledge that dependence between T_1 and T_2 is likely complex

- The illness-death model that has been presented so far, that is:

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, && t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, && t_2 > 0 \\
h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2|t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, && t_2 > t_1
\end{aligned}
$$

  has two components that 'represent' dependence

- One is the shared subject-specific frailty, γ_i
  - discussed above

- The second is in the interplay between the two hazards for the terminal event
  - prior to the occurrence of the non-terminal event, $h_2(t_2)$
  - following the occurrence of the non-terminal event, $h_3(t_2|t_1)$

- One approach to summarizing this interplay is the *explanatory hazard ratio*:

$$\text{EHR}(t_1, t_2) = \frac{h_3(t_2|t_1)}{h_2(t_2)} = \frac{h_{03}(t_2|t_1)\ \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}}{h_{02}(t_2)\ \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}}$$

  - does not depend on the value of γ_i
  - nevertheless, a reasonably complex function

- Work needs to be done to develop best practices around reporting and interpreting $\text{EHR}(t_1, t_2)$
  - e.g. graphical strategies for reporting

- In the meantime, one take-away from $\text{EHR}(t_1, t_2)$ is that the form of dependence between T_1 and T_2 is reasonably flexible

Markov illness-death models

- As mentioned, $h_{03}(t_2|t_1)$ has the potential to be quite complex
  - a continuous function, $h_{03}(t_2|\cdot)$, that is a function of a continuous t_1
- One way forward is the so-called *Markov illness-death model*:

$$\begin{aligned} h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, & t_1 > 0 \\ h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, & t_2 > 0 \\ h_3(t_2|t_1) &= \gamma_i \; h_{03}(t_2) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, & t_2 > t_1 \end{aligned}$$

- Intuition underpinning this choice:
  - $h_2(t_2)$ is the hazard that initially dictates risk for the terminal event
  - if/when a patient experiences the non-terminal event, risk is subsequently dictated by $h_3(t_2)$

- The Markov assumption amounts to 'forgetting' when the non-terminal event occurred

- An even simpler Markov illness-death model is:

$$h_1(t_1) = \gamma_i \, h_{01}(t_1) \, \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, \qquad t_1 > 0$$
$$h_2^*(t_2) = \gamma_i \, h_{02}^*(t_2) \, \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2^* + Z_i(t_2)\beta_z\}, \qquad t_2 > 0$$

  where

$$Z_i(t_2) = \begin{cases} 0 & \text{if the non-terminal event has not occurred by } t_2 \\ 1 & \text{if the non-terminal event has occurred by } t_2 \end{cases}$$

  * treat the non-terminal event as a time-varying covariate in a single hazard model for mortality

Semi-Markov illness-death model

- An alternative is to re-orient the 3rd transition so that the focus is on modeling the *sojourn time*
  - i.e. $T_2 - T_1$
  - e.g. time to delivery following a diagnosis of preeclampsia

$$
\begin{aligned}
h_1(t_1) &= \gamma_i \; h_{01}(t_1) \; \exp\{\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_1\}, & t_1 > 0 \\
h_2(t_2) &= \gamma_i \; h_{02}(t_2) \; \exp\{\boldsymbol{X}_{2i}^T \boldsymbol{\beta}_2\}, & t_2 > 0 \\
h_3(t_2 - t_1) &= \gamma_i \; h_{03}(t_2 - t_1) \; \exp\{\boldsymbol{X}_{3i}^T \boldsymbol{\beta}_3\}, & t_2 > t_1
\end{aligned}
$$

- Referred to as a semi-Markov model for the baseline hazard
  - shift of time scale together with 'forgetting' when the shift occurred
- Fundamental change in the time scale for $h_3(\cdot)$
  - post-preeclampsia person-time is re-aligned

- Corresponding fundamental change in the interpretation of $\boldsymbol{\beta}_3$
  - arguably should not compare results with those from a Markov model

- As a modeling choice, one could consider including the timing of the non-terminal event (i.e. T_{1i}) in $\boldsymbol{X}_{3i}$
  - consider whether the timing of delivery following a diagnosis of preeclampsia depends on when the diagnosis occurred

Joint risk prediction

- At any given point in time, a patient could be in one of four 'outcome categories'

- Conceive of a patients *joint risk profile* consisting of four probabilities

$$\begin{aligned}
\pi^{(1)}(t) &= \Pr(\text{non-terminal alone at time } t) \\
\pi^{(2)}(t) &= \Pr(\text{both at time } t) \\
\pi^{(3)}(t) &= \Pr(\text{terminal alone at time } t) \\
\pi^{(4)}(t) &= \Pr(\text{neither at time } t)
\end{aligned}$$

  * probabilities add up to 1.0

- Can be calculated following the fit of an illness-death model
  * Putter et al (*Statistics in Medicine*, 2007)

- Furthermore, the predicted risk profiles can be calculated:
  - to be covariate-specific
  - at a range of relevant time points

- Such individualized predictions have the potential to be useful in the course of patient/clinician decision-making
  - Reeder et al (*Circulation: CQO*, 2019)

Software

- Various options that differ in their functionality
- We have been developing the `SemiCompRisks` package for `R`
  - Alvarez et al (*R Journal*, 2019)
- Broad, flexible functionality:
  - data type
    - independent and cluster-correlated
  - baseline hazard specifications
    - parametric and semi-parametric
  - estimation and inference
    - frequentist and Bayesian paradigms
  - regression modeling framework
    - hazard-based and accelerated failure time models

- Other, relevant packages for `R`
  - `frailtypack` for `R`
  - `mstate` for `R`
  - `multistate` for `STATA`

Example using data from BIDMC

- In this section, we present a worked analysis using data from the electronic health record (EHR) of Beth Israel Deaconess Medical Center (BIDMC)
- Goal is to illustrate process, considerations, practical details, and limitations
- Specific topics:
  - data description, manipulation, and checks
  - exploratory data analysis
  - univariate modeling of covariate/outcome relationship
  - illness-death modeling
  - individualized profile prediction
- Minimal code will be incorporated in the slides
  - a complete analogous example using simulated data is available

Data description

- Data on N=5,054 singleton deliveries in 2016
- Demographic characteristics recorded at patient intake
  - e.g. age, race, insurance status
- Lab values measured during 42 weeks preceding delivery
  - values annotated with 'abnormal' binary indicator
- ICD-10 diagnostic codes associated with delivery
  - pregnancies with preeclampsia defined by corresponding codes
  - also identifies certain baseline chronic conditions
    - e.g. anemia, diabetes, obesity

Outcomes

- 319/5,054 (6.3%) of patients diagnosed with preeclampsia during pregnancy
- Unfortunately, because there was no systematic screening, we do not have the exact date of the onset of preeclampsia
- The EHR does, however, have information on:
  - date of hospital admission
  - date of delivery
  - gestational age at delivery
- We use hospital admission time as proxy
  - approach used by other studies of preeclampsia risk
  - e.g. Lisonkova and Joseph (*AJOG*, 2013)

- Furthermore, we use this information to derive admission and delivery timing on gestational age (GA) scale

- Because preeclampsia definitionally cannot onset before 20 weeks, we rescale times to (GA - 20) weeks
  - i.e. set the 'origin' to be 20 weeks gestation

- Finally, for preeclampsia patients admitted and delivering same day, we must round sojourn time up to ≥ 1 day

Data cleaning and formatting

- Data cleaning is always project-specific, but the setting of an EHR-based study of preeclampsia poses particular challenges worth highlighting

- Dataset actually includes all births from 2000-2016
  - multi-year dataset with multiple pregnancies for some mothers
  - simplified by restriction to 2016 births

- Patients have different measurements taken at different times throughout pregnancy
  - only consider those that are well-defined at 'baseline'
    - i.e. at the 20-week mark of the pregnancy
  - absence of certain measurements is taken as indicator of 'normal' status
    - need to consider whether this has the potential to result in bias

- On a technical note, for `SemiCompRisks` package, the data must be in 'wide' format with one patient per row

ID	Age	...	Y1	delta1	Y2	delta2
001	28	...	39	0	39	1
002	32	...	42	0	42	1
003	23	...	32	0	32	1
004	37	...	29	1	39	1
005	34	...	33	1	34	1
006	19	...	39	0	37	1
007	41	...	39	1	40	1
...						

- Also:
  - when $\delta_{i1} = 0$, must have $y_{i1} = y_{i2}$
  - when $\delta_{i1} = 1$, must have $y_{i1} < y_{i2}$

Summary of covariates: demographics

- Based on patient intake forms
  - constrained by categories captured in EHR

	Both	Delivery only	p-value
	n=319	n=4735	
Age $<$ 30	103 (32.3%)	1192 (25.2%)	$<$0.001
Age 30 to 35	102 (32.0%)	2098 (44.3%)	
Age 35 and up	114 (35.7%)	1445 (30.5%)	
Private insurance	254 (79.6%)	3914 (82.7%)	0.171
White	143 (44.8%)	2026 (42.8%)	$<$0.001
Black	60 (18.8%)	504 (10.6%)	
Hispanic	13 (4.1%)	229 (4.8%)	
Asian	19 (6.0%)	726 (15.3%)	
Other or unknown race	84 (26.3%)	1250 (26.4%)	

Summary of covariates: medical history

- Defined by ICD-10 codes associated with delivery
- Coding:
  - absence of code taken as absence of condition
  - history of preeclampsia is defined by the presence of a preeclampsia diagnosis at a previous delivery at BIDMC

	Both	Delivery only	p-value
	n=319	n=4735	
Previous PE	18 (5.6%)	58 (1.2%)	<0.001
BMI>30	58 (18.2%)	234 (4.9%)	<0.001
Anemia	32 (10.0%)	217 (4.6%)	<0.001
Diabetes	20 (6.3%)	76 (1.6%)	<0.001
Parity 0	189 (59.2%)	2327 (49.1%)	0.002
Parity 1	93 (29.2%)	1661 (35.1%)	
Parity 2+	37 (11.6%)	747 (15.8%)	

Summary of covariates: lab values

- Uses latest labs recorded by week 20
  - absence of 'abnormal' lab taken equivalent to observed 'normal' lab

	Both	Delivery only	p-value
	n=319	n=4735	
Abnormal platelet count	13 (4.1%)	62 (1.3%)	0.001
Abnormal white blood cell count	57 (17.9%)	548 (11.6%)	0.002
Abnormal hematocrit	30 (9.4%)	332 (7.0%)	0.116
Abnormal hemoglobin	28 (8.8%)	276 (5.8%)	0.038

- Summary tables highlight high-level relation between baseline covariates and outcome events

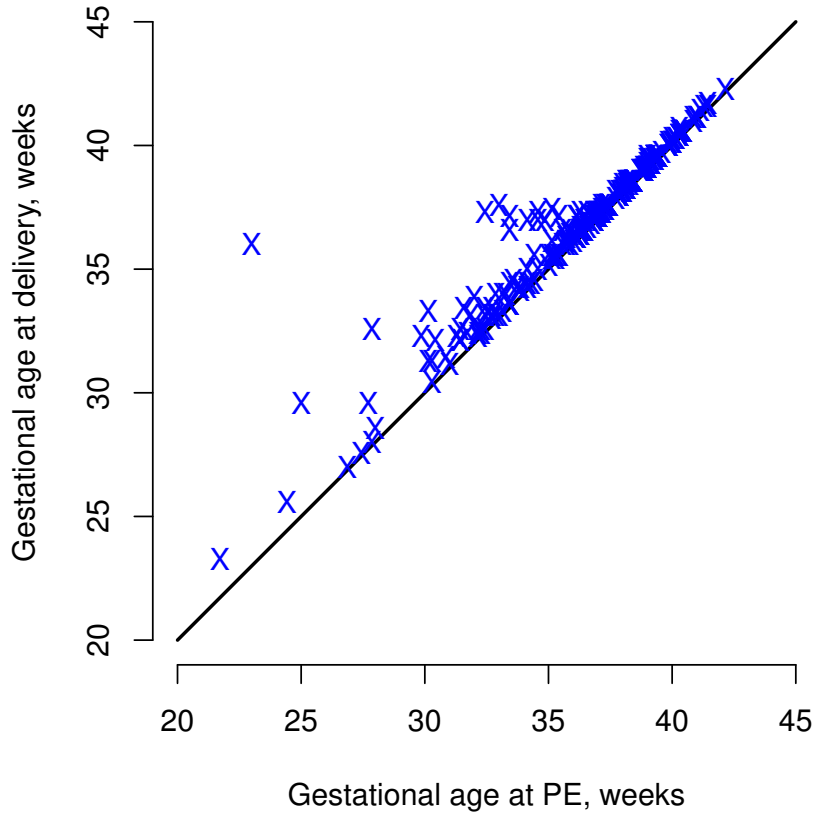
- In this setting, since patients deliver (either with or without preeclampsia) there is no censoring
  - reduction of four typical observed categories of neither, non-terminal only, terminal only, or both

- Summary tables, however, lack insight into event *timing*

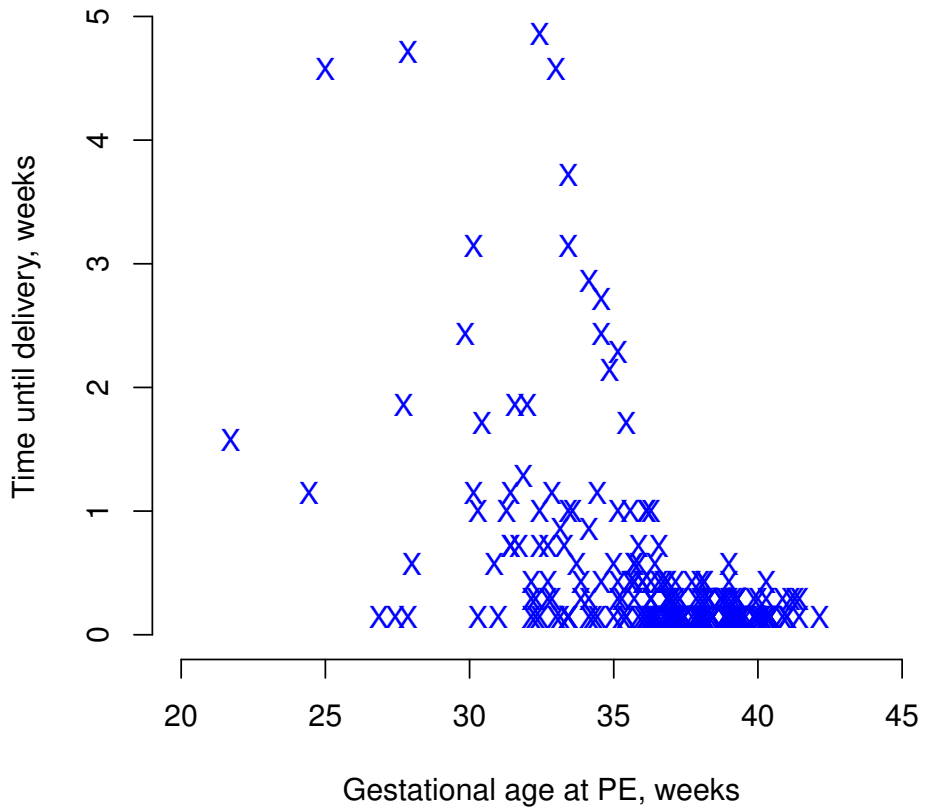
- Additional exploratory data analysis (EDA) should be useful

EDA for the outcomes among women diagnosed with preeclampsia

- Both on the 'gestational age' time scale



- Also, examine the sojourn time:



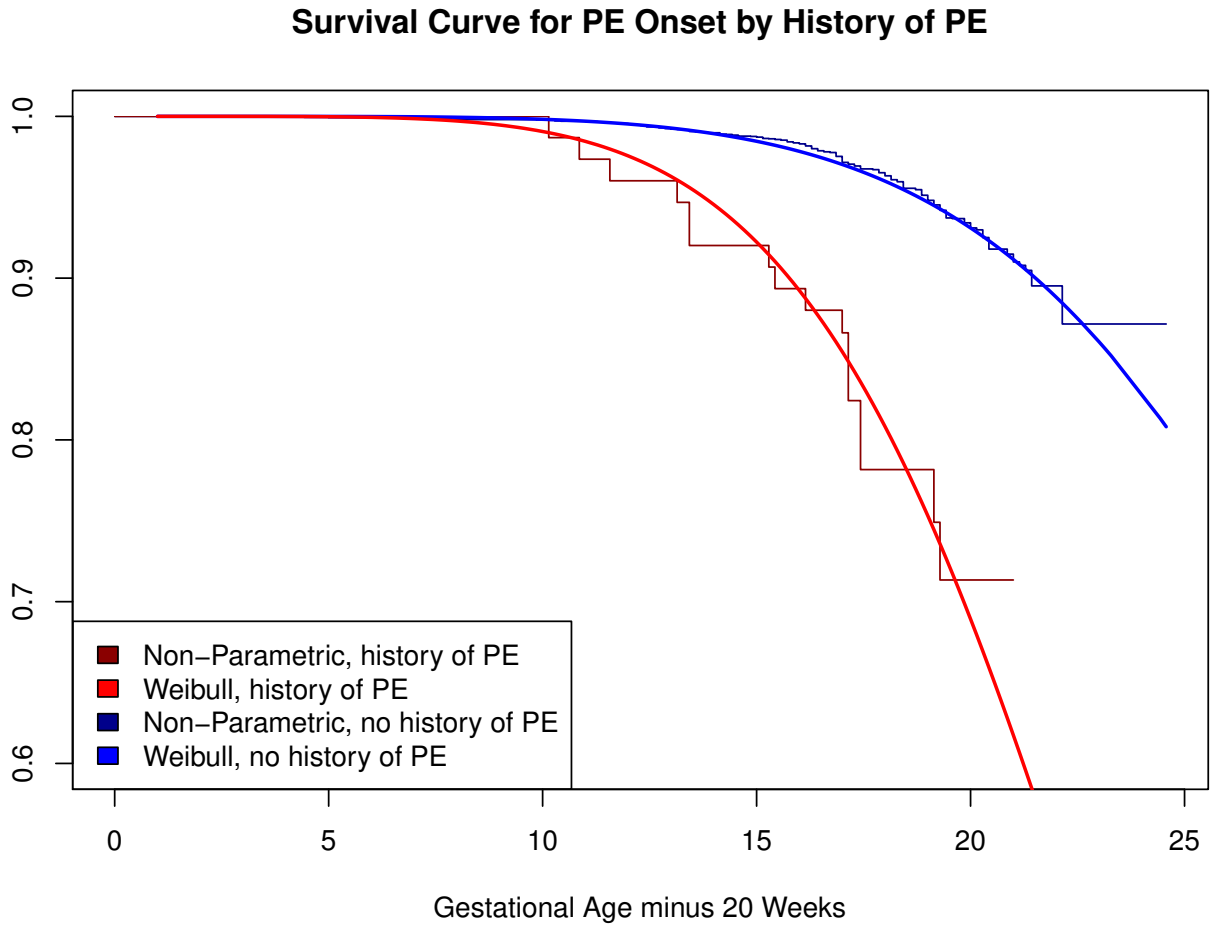
- Clear trend in preeclampsia timing throughout course of pregnancy
  - time from preeclampsia onset to delivery short for late-onset preeclampsia
  - can be longer for early-onset preeclampsia

- May reflect clinical decision-making process
  - balance risks from preterm delivery with risks from prolonging pregnancy with preeclampsia

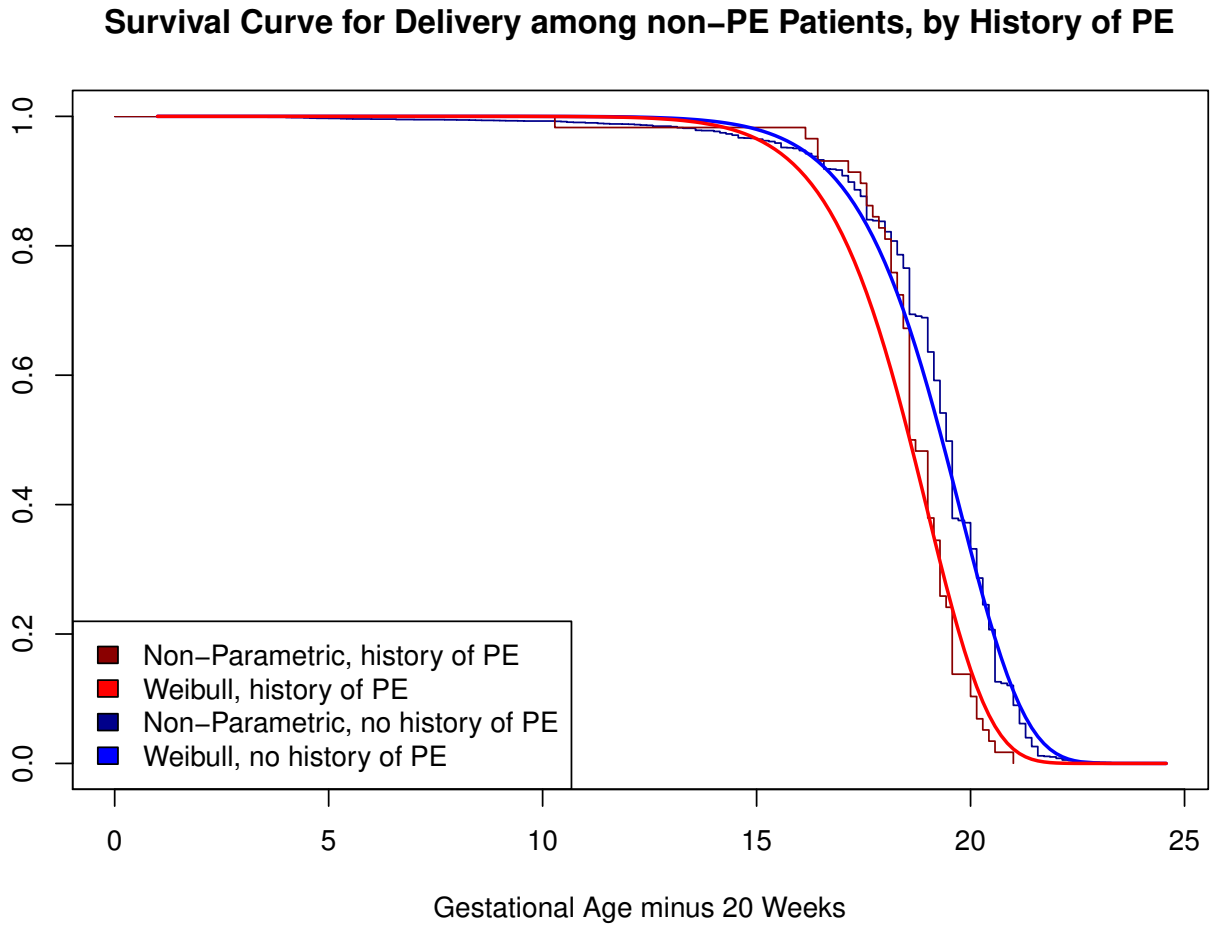
- Observed structure also informs modeling decisions, particularly transition assumption
  - Markov assumption that delivery timing does not directly depend on time of preeclampsia onset may not be reasonable
  - a semi-Markov assumption for $h_3(t_2|t_1)$ with T_1 as a covariate seems more reasonable

Univariate models

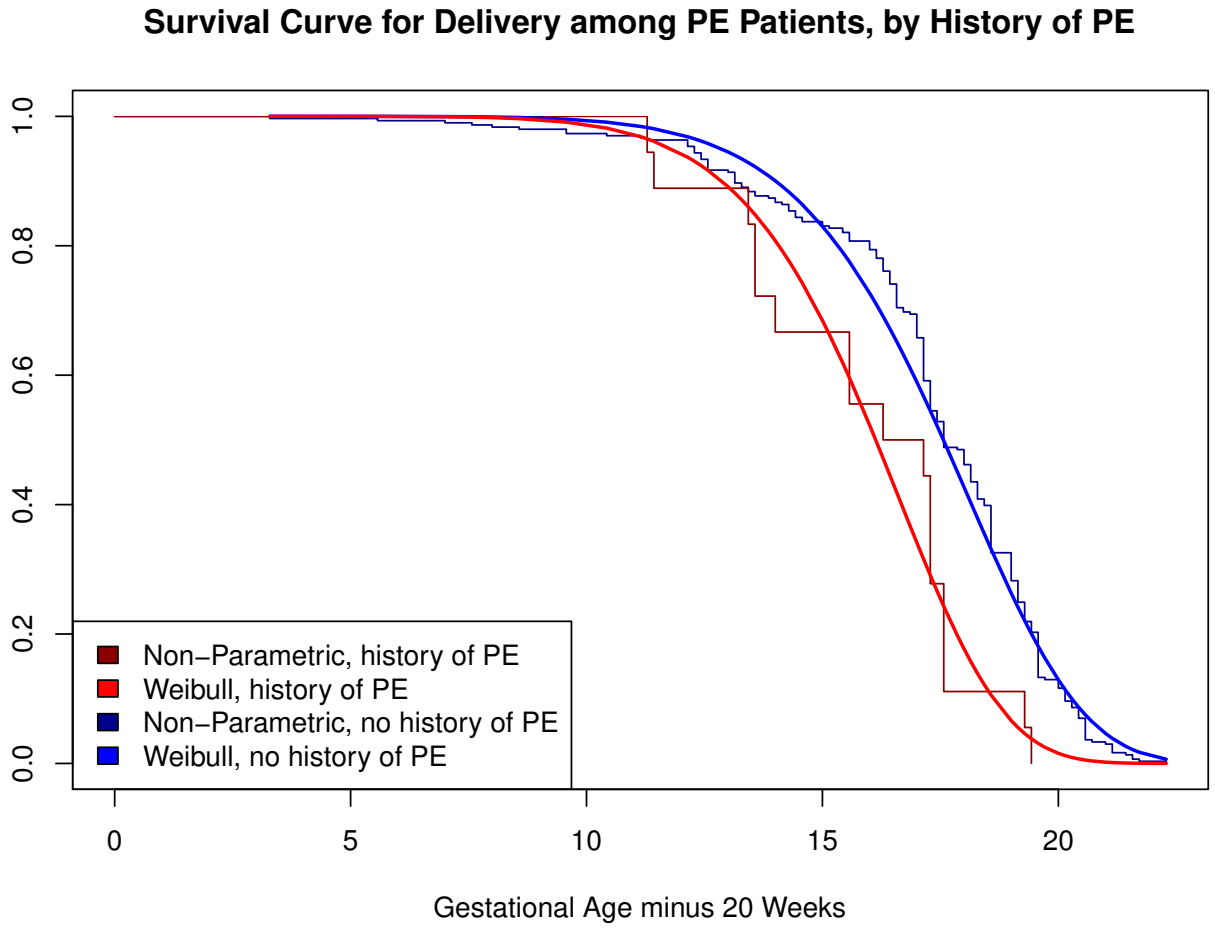
- Before building joint models, univariate models and plots can help illustrate potential covariate-outcome dynamics
- For each transition submodel, univariate exploration allows us to:
  - assess strength and direction of univariate association with time-to-event
  - assess fit of parametric model specification
- Here we look at examples of such EDA for a particular covariate
  - parametric Weibull model fit using `flexsurv` package
  - results plotted with `plot` function
- Event times further scaled (GA-20)/10 going forward
  - putting baseline parameters and regression coefficients on similar scale
  - helps numerically with the mechanics of model fitting algorithms



- Estimated hazard ratio:
  - 5.21 based on a Weibull model fit
  - 5.33 based on a standard Cox model fit



- Estimated hazard ratio:
  - 1.74 based on a Weibull model fit
  - 1.95 based on a standard Cox model fit



- Estimated hazard ratio:
  - 2.03 based on a Weibull model fit
  - 2.13 based on a standard Cox model fit

- Examples show associations consistent with scientific understanding
  - i.e. a history of preeclampsia is associated with increased risk in the current pregnancy
- Weibull parametric survival curves are close but not identical to non-parametric survival curves
  - covariate effect estimates similar in parametric vs. semi-parametric Cox analyses
- There is some evidence that the effect of a history of preeclampsia on $h_2(\cdot)$ and $h_3(\cdot)$ is not constant over time
  - i.e. non-proportional hazards
  - caution against over-interpreting univariate models, as conditioning on other covariates, incorporating frailties, etc introduces flexibility
  - could consider alternative model specification, such as AFT illness-death model

Illness-death modeling

- Example will focus on an illness-death model with
  - a semi-Markov specification for $h_3(t_2|t_1)$
  - Weibull baseline hazard functions

- Fit using the `FreqID_HReg()` function in `SemiCompRisks` package

```
FreqID_HReg(Formula,
            data,
            model="semi-Markov",
            frailty=TRUE, ...)
```

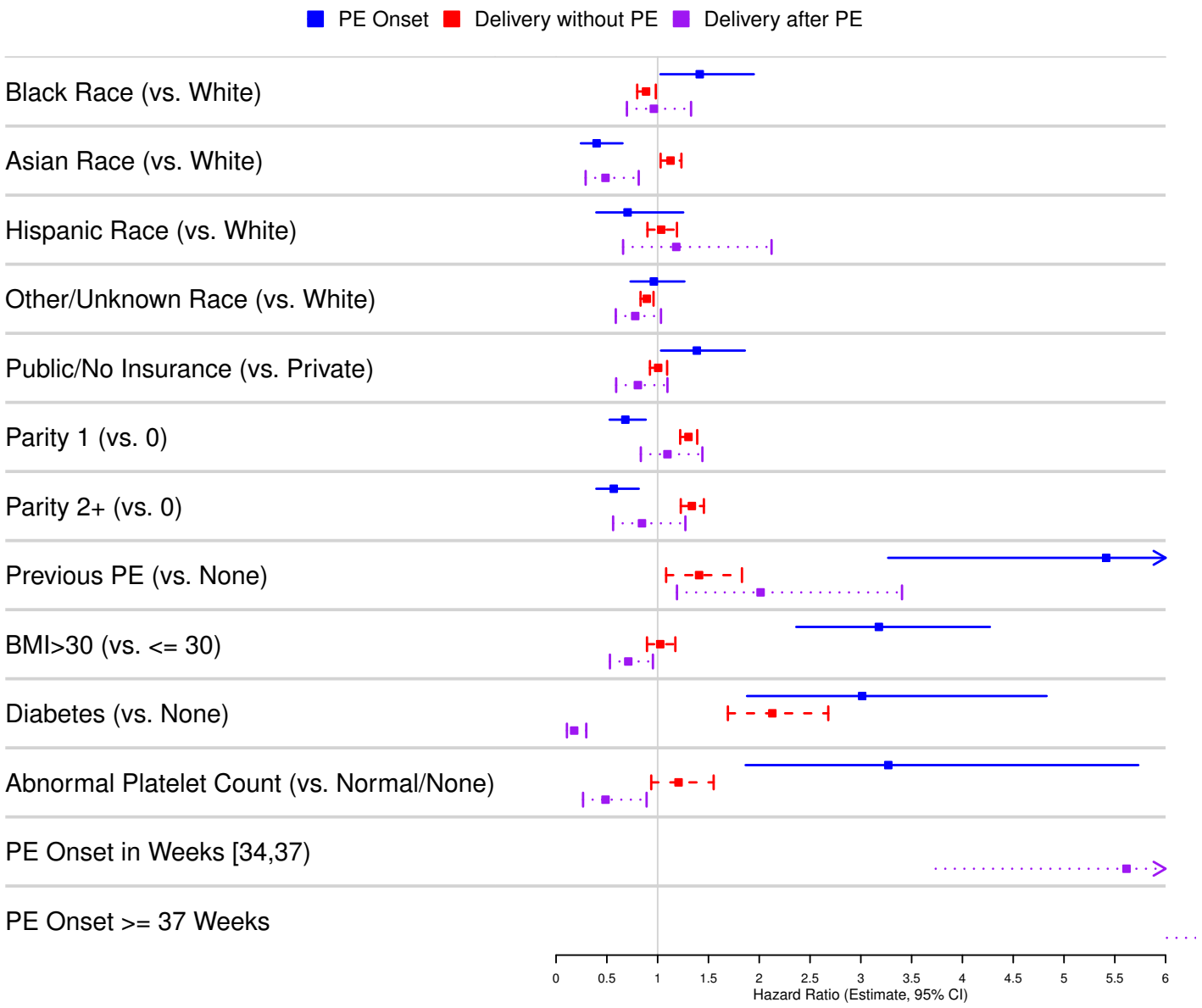
  - `Formula` takes the form

    `y1 + delta1 | y2 + delta2 ~ x1 + x2 | x1 + x2 | x1 + x2`

    - uses variable names from `data`
  - `model` specifies either Markov or semi-Markov
  - `frailty` specifies inclusion of Gamma distributed shared frailty

- As evidenced by syntax, the illness-death model admits covariates in all three transition submodels
  - simplest to specify same covariates across all hazards
  - though loose efficiency if a covariate that is not associated with risk is included
  - data-driven variable selection methods are an area of active research (ours!)
- In this example, we specify same full model across all hazards based on prior knowledge and EDA
- Additionally include a categorical effect of T_1 on $[20, 34]$, $(34, 37]$, and $[37, 45)$ in the model for the sojourn time

Results: covariate effects



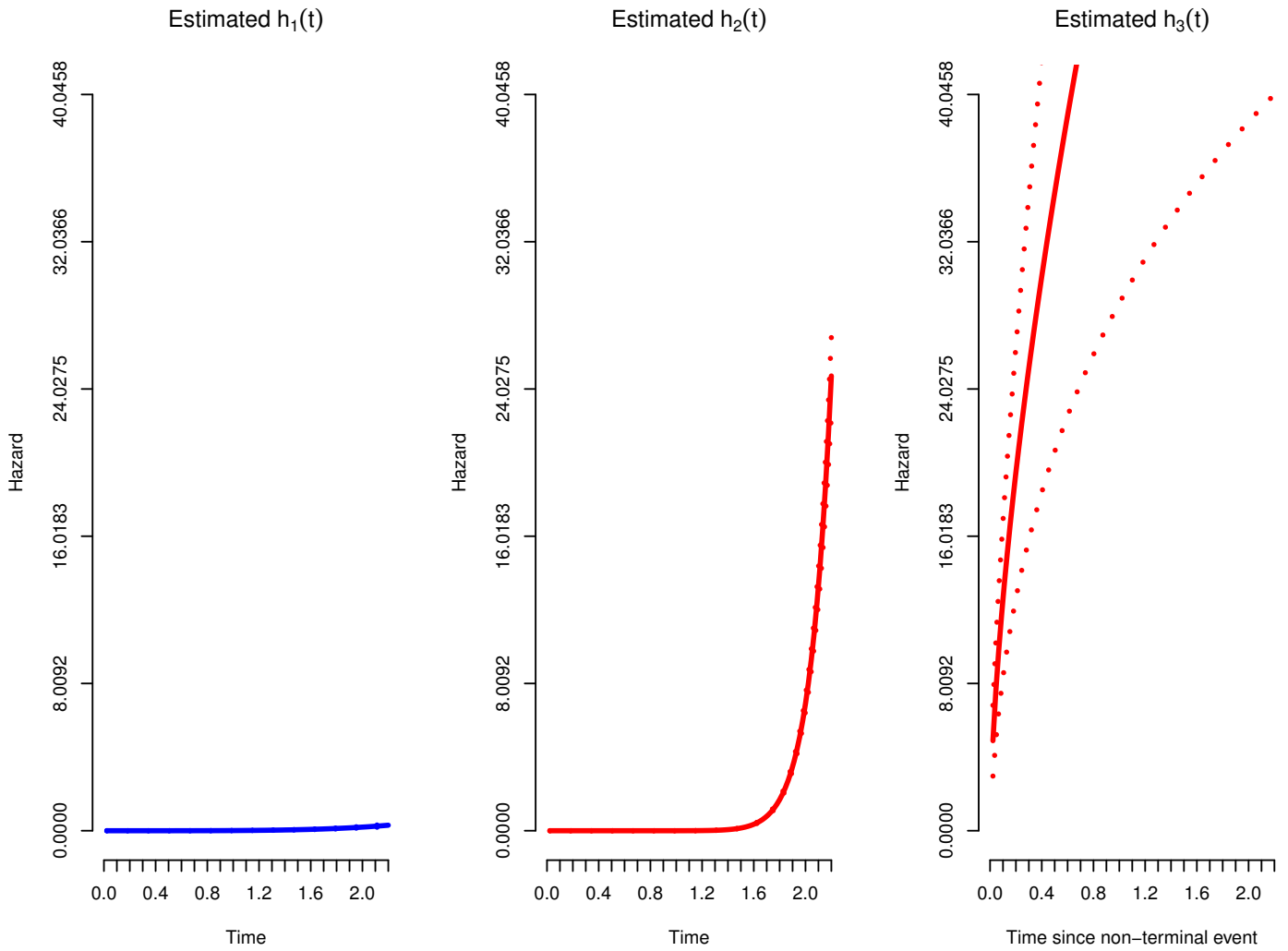
- Each value represents hazard ratio for the corresponding transition hazard, *fixing all other covariates and the frailty*

- Overall, analysis indicates that conditional on other covariates, conditions like anemia and elevated BMI are associated with earlier preeclampsia onset, and thus are also associated with longer time from preeclampsia onset to delivery

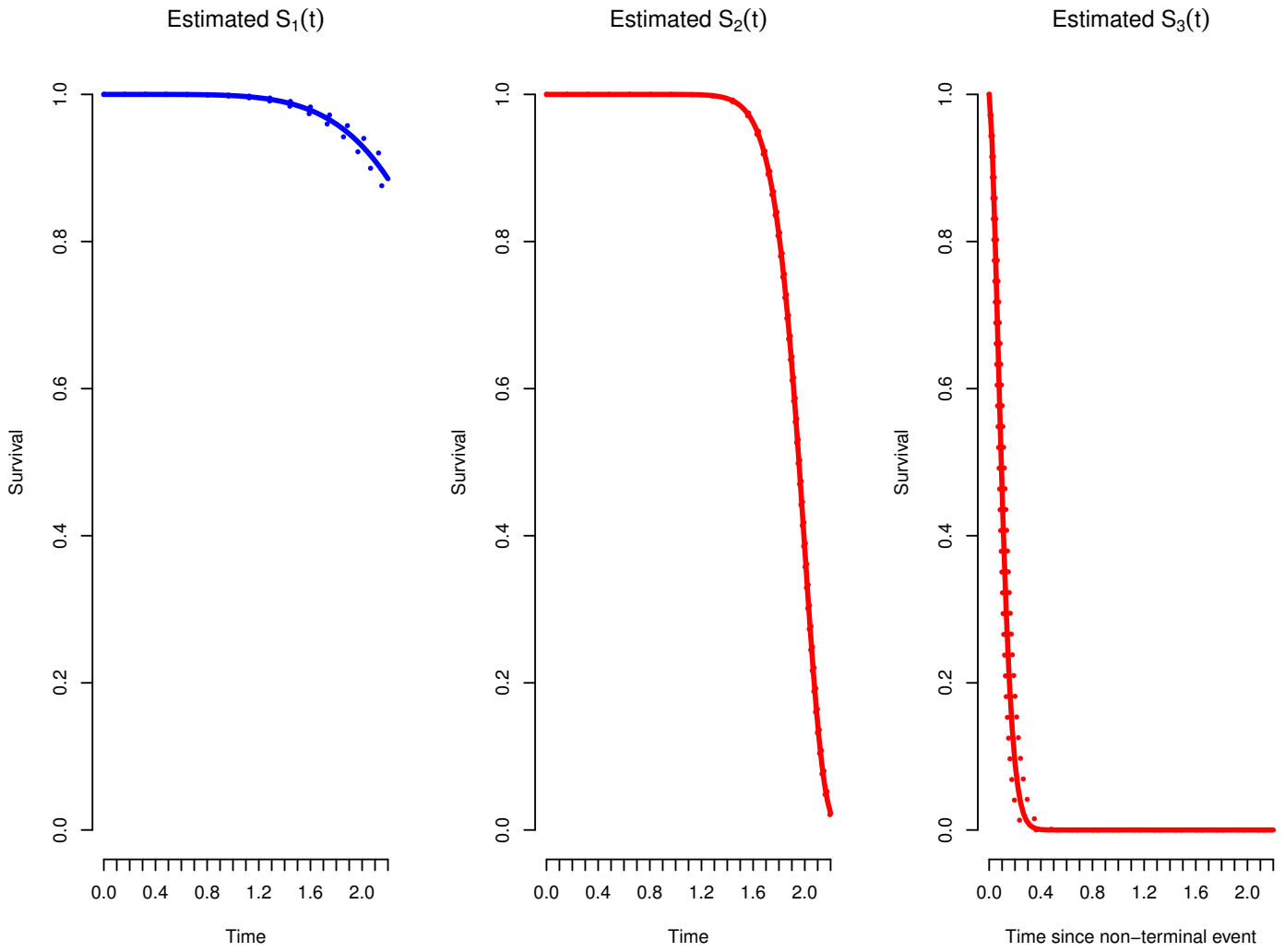
- Estimated frailty variance is $\hat{\theta}$=0.006, with a 95% CI of (0.00003, 0.046)
  - quite small!
  - indicates that covariates and model structure account for much of the outcome dependence

- Using `plot` and `predict` functions, can also examine the estimated hazard and survivor functions fixed covariates

Results: baseline hazard functions



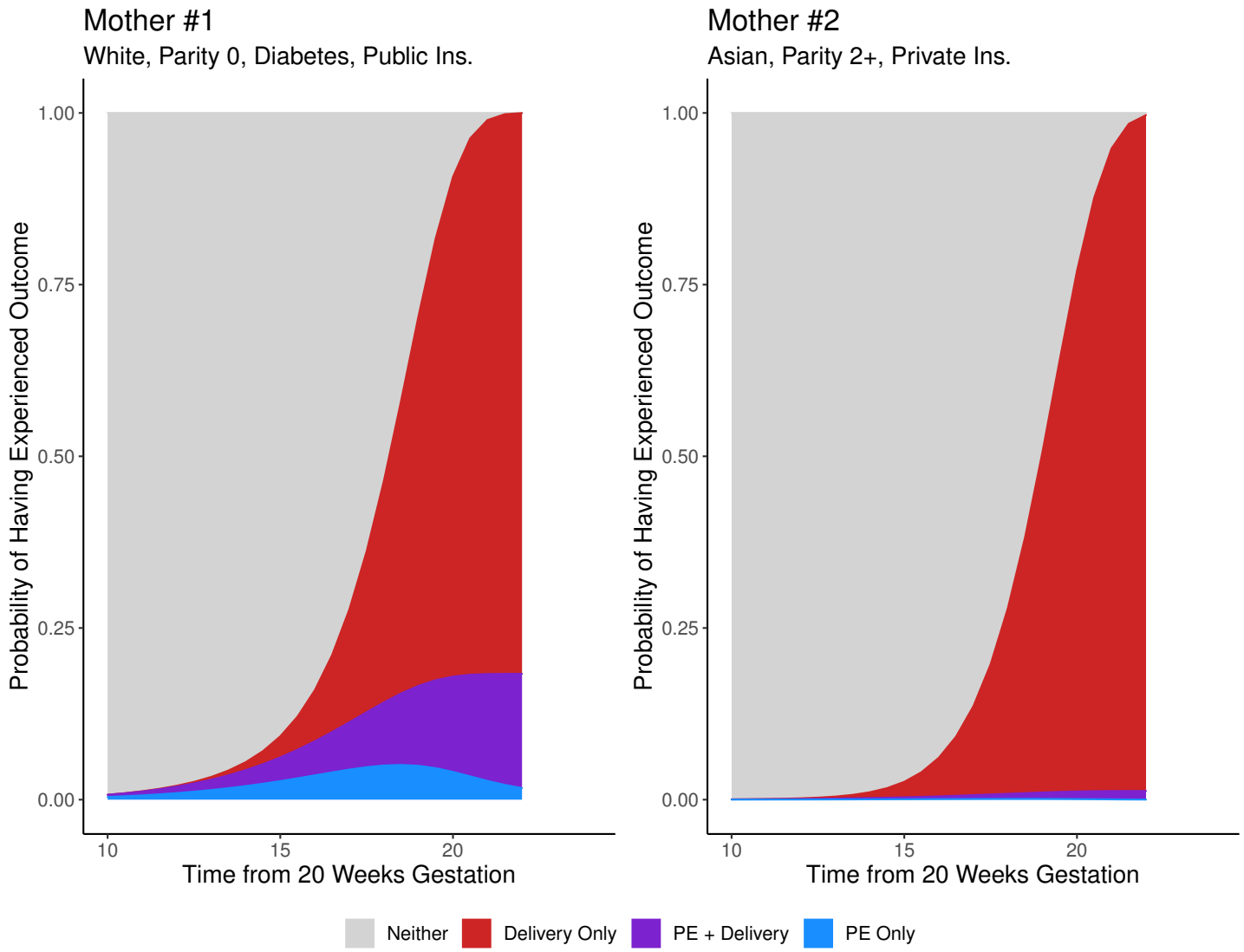
Results: baseline survivor functions



- Estimated hazard and survivor function plots illustrate useful curves for each transition hazard

- Lastly, we might use predicted risk profile to further summarize the joint outcome trajectory of a particular patient
  - showing across future time points the absolute risk of being in each outcome category
  - offers individualized insight into predicted outcomes, with value for patient/clinician decision-making

Results: predicted risk profiles for two mothers



Final comments

- Semi-competing risks arise in a broad range of settings and yet are under-appreciated
- Provide an opportunity to frame research questions in a way that may yield new insight
  - acknowledge, exploit and investigate dependence
- Motivated by thinking through issues in the study of preeclampsia
- Our sense is that there substantial promise for these methods in pediatric and perinatal research
  - in relation to conditions that arise during pregnancy
  - when the force of mortality is relatively strong
- `SemiCompRisks` package for `R`

Additional topics (not covered)

- Accelerated failure time models
  - an alternative to models for the hazard function
  - interpret covariate effects in terms of the impact on median survival
  - Lee et al (*Biometrics*, 2017)

- Cluster-correlated data
  - e.g. patients within hospitals
  - flexible Bayesian framework for estimation/inference
  - Lee et al (*JASA*, 2016)

- Outcome-dependent sampling
  - analysis of semi-competing risks data from a nested case-control study
  - novel design for sub-sampling on the basis of the non-terminal and terminal events simultaneously
  - Jazic et al (*Statistical Methods in Medical Research*, 2020)

- Time-to-event analysis when the timescale is finite
  - time-to-event analysis typically assume that, absent competing risks, a patient will eventually experience the event
  - not always the case
  - e.g. preeclampsia
  - Lee et al (*Statistical Methods in Medical Research*, 2020)

- Variable/feature selection in high-dimensional settings
  - regularization methods based on carefully chosen penalties
  - regularize within and between the three hazard functions
  - Reeder et al (soon to appear on *arXiv*)

References

Preeclampsia

- Lisonkova, Sarka, and K. S. Joseph. Incidence of preeclampsia: risk factors and outcomes associated with early-versus late-onset disease. American Journal of Obstetrics and Gynecology 209.6 (2013): 544-e1.
- Townsend, R., Khalil, A., Premakumar, Y., Allotey, J., Snell, K.I., Chan, C., Chappell, L.C., Hooper, R., Green, M., Mol, B.W. and Thilaganathan, B., 2019. Prediction of pre-eclampsia: review of reviews. Ultrasound in Obstetrics & Gynecology, 54(1), pp.16-27.
- Giannakou, K., Evangelou, E., & Papatheodorou, S. I. (2018). Genetic and non-genetic risk factors for pre-eclampsia: umbrella review of systematic reviews and meta-analyses of observational studies. Ultrasound in Obstetrics & Gynecology, 51(6), 720-730.

Non-statistical papers

- Haneuse S, Lee KH. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is non-terminal. Circulation: Cardiovascular Quality and Outcomes. 2016; 9:322-331.
- Jazic I, Schrag D, Sargent D, Haneuse S. Beyond composite endpoints analysis: semi-competing risks as an underutilized framework for cancer research. Journal of the National Cancer Institute. 2016; 108(12).
- Reeder H, Shen C, Buxton E., Haneuse S, Kramer D. Joint shock/death risk prediction model for patients considering implantable cardioverter-defibrillators: A secondary analysis of the SCD-HeFT trial. Circulation: Cardiovascular Quality & Outcomes. 2019; 12(8): e005675.
- Crilly C, Haneuse S, Litt J. Predicting the outcomes of preterm neonates beyond the neonatal intensive care unit: What are we missing? Pediatric Research. In press.

Causal inference

- Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by 'death'. Journal of Educational and Behavioral Statistics 28(4), 353-368.
- Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. Biometrics. 2005;61:305-310.
- Mattei A, Mealli F. Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. Biometrics. 2007;63:437-446.
- Jemiai Y, Rotnitzky A, Shepherd BE, Gilbert PB. Semiparametric estimation of treatment effects given baseline covariates on an outcome measured after a post-randomization event occurs. JRSS-B. 2007;69:879-901.
- Egleston, B. L., Scharfstein, D. O., Freeman, E. E., West, S. K. (2007). Causal inference for non-mortality outcomes in the presence of death. Biostatistics, 8(3), 526-545.

- Chiba Y, Vanderweele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. Am Journal Epidemiology. 2011;173:745-751.

- Tchetgen Tchetgen, E. J. (2014). Identification and estimation of survivor average causal effects. Statistics in Medicine 33(21), 3601-3628

Copula-based methods

- Fine JP, Jiang H, Chappell R. On semi-competing risks data. Biometrika. 2001;88:907-919.
- Peng, L. and J. P. Fine (2007). Regression modeling of semicompeting risks data. Biometrics 63(1), 96-108.
- Hsieh, J.-J., W. Wang, and A. Adam Ding (2008). Regression analysis based on semi-competing risks data. JRSS-B 70(1), 3-20.

Illness-death models

- Putter, H., Fiocco, M., & Geskus, R. (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine, 26(11), 2389-2430.
- Kneib, T. and A. Hennerfeind (2008). Bayesian semiparametric multi-state models. Statistical Modeling 8, 169-198.
- Xu, J., J. D. Kalbfleisch, and B. Tai (2010). Statistical analysis of illness-death processes and semicompeting risks data. Biometrics 66(3), 716-725.
- Zeng, D., Q. Chen, M.-H. Chen, J. G. Ibrahim, et al. (2012). Estimating treatment effects with treatment switching via semicompeting risks models: an application to a colorectal cancer study. Biometrika 99(1), 167-184.
- Lee KH, Haneuse S, Schrag D, Dominici F. Bayesian semi-parametric analysis of semi-competing risks data: Investigating hospital readmission after a pancreatic cancer diagnosis. JRSS-C 2015;64(2):253.

- Lee KH, Dominici F, Schrag D, Haneuse S. Hierarchical models for semi-competing risks data with application to quality of end-of-life care for pancreatic cancer. Journal of the American Statistical Association. 2016; 111(515):1075-1095.

- Lee KH, Rondeau V, Haneuse S. Accelerated failure time models for semi-competing risks data in the presence of complex censoring. Biometrics. 2017;73(4):1401-1412.

- Alvarez D, Haneuse S, Lee C, Lee KH. SemiCompRisks: An R Package for Independent and Cluster-Correlated Analyses of Semi-Competing Risks Data. The R Journal. 2019. Vol 9/1.

- Jazic I, Lee S, Haneuse S. Estimation and inference for semi-competing risks based on data from a nested case-control study. Statistical Methods in Medical Research. In press.

- Lee C, Lee S, Haneuse S. Time-to-event analysis when the event is defined on a finite time interval. Statistical Methods in Medical Research. In press.