Super Learning for Combining Dynamic Predictions for Decision-Making

Dimitris Rizopoulos[1] **and Jeremy M.G. Taylor**[2]

[1]Department of Biostatistics, Erasmus Medical Center Rotterdam
[2]Department of Biostatistics, University of Michigan

d.rizopoulos@erasmusmc.nl
jmgt@umich.edu

@drizopoulos

Super Learning for Dynamic Predictions

1 Background & Motivation

Setting: Follow-up studies

- ▷ multiple longitudinal outcomes
  - * biomarkers
  - * patient parameters
  - * patient reported outcome scores

- ▷ one or multiple endpoints
  - * relapse of disease
  - * requirement for intervention
  - * death

1 Background & Motivation (cont'd)

Obtain accurate predictions for the (cumulative) risk of an event to guide decision making

Using the available longitudinal information

1 Background & Motivation (cont'd)

Erasmus MC
Universitair Medisch Centrum Rotterdam

University of Michigan Prostatectomy Data

- 3634 PCa patients followed-up in 1996–2013
  - aged 40 to 84 years with clinically localized cT1 to cT3 disease
  - received radical prostatectomy

- We excluded patients who
  - had Gleason score ≤ 4
  - initiated any ADT more than 1 year before treatment

- baseline variables: PSA, Gleason, T-stage, age, race, gland volume, perineural invasion, planned adjuvant therapy

1 Background & Aim (cont'd)

University of Michigan Prostatectomy Data

Patients remain at risk of metastasis

- Follow-up
  - * PSA levels at frequent intervals
  - * when PSA increases, physicians consider Salvage Therapy (ST)
  - * ST androgen deprivation therapy, radiation therapy, chemotherapy, and combinations

1 Background & Motivation (cont'd)

University of Michigan Prostatectomy Data

Use the longitudinal PSA & baseline covariates to predict the risk of metastasis

1 Background & Motivation (cont'd)

- Two main frameworks to obtain such predictions

  ▷ *Landmarking*

  * a series of Cox models at different follow-up times
  * last value of the biomarker as a baseline covariate
  * Breslow estimator of survival probabilities

  ▷ *Joint Models*

  * complete specification of the joint distribution of the outcomes
  * direct derivation of conditional risk probabilities

1 Background & Motivation (cont'd)

Landmarking

- ▷ *Advantages*
  - * easier to use, available in standard software
  - * can generalize to multiple biomarkers without (much) extra computational cost

- ▷ *Disadvantages*
  - * predictions not consistent
  - * not plausible LOCF for biomarkers
  - * does not account for measurement error and endogeneity
  - * not valid causal interpretation

1 Background & Motivation (cont'd)

Joint Models

- ▷ *Advantages*
  - * consistent predictions
  - * accounts for measurement error and endogeneity
  - * biomarkers follow a trajectory
  - * valid causal interpretation

- ▷ *Disadvantages*
  - * computationally intensive
  - * *sensitive to modeling assumptions*

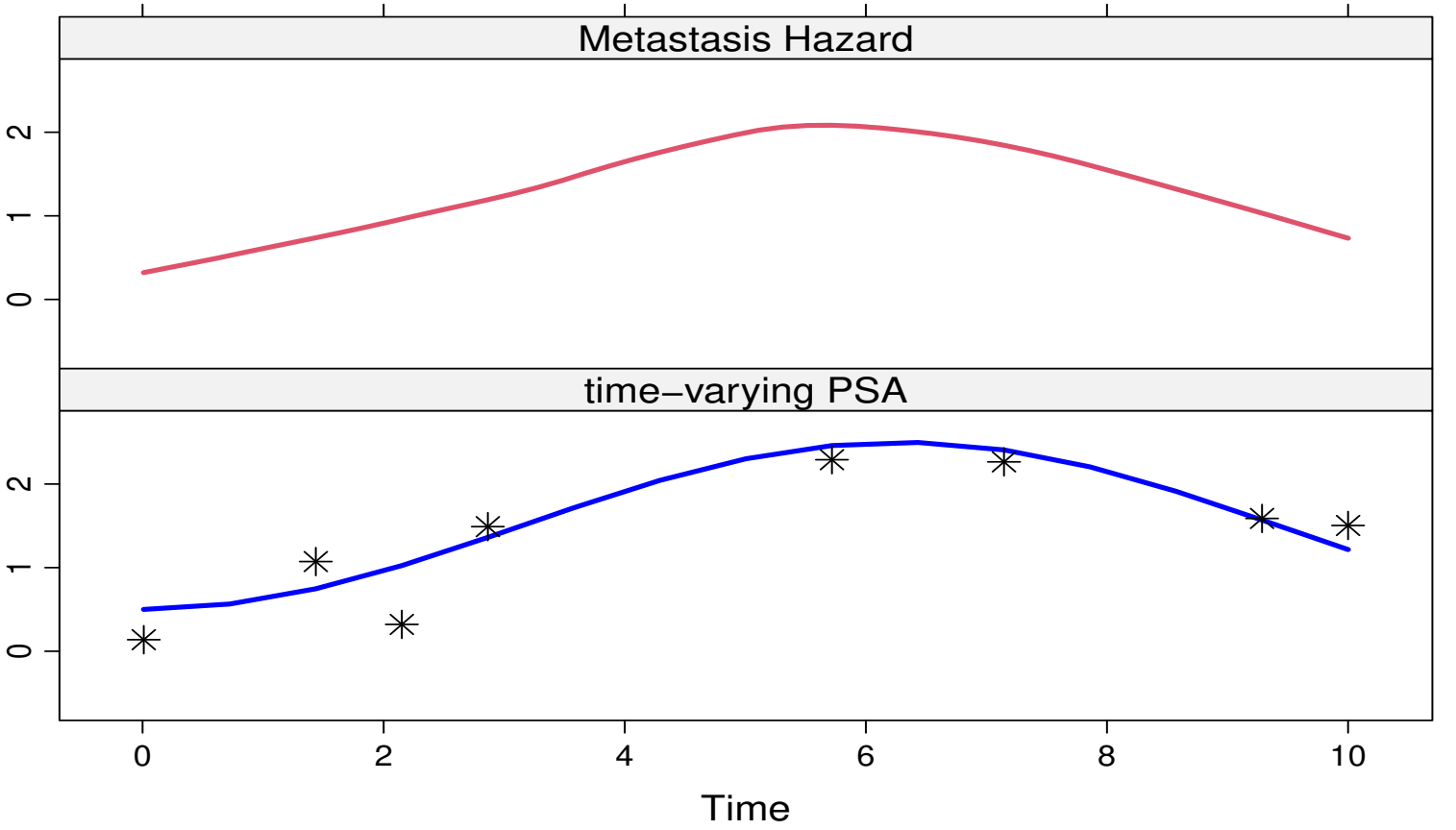
1 Background & Motivation (cont'd)

- *Sensitive to modeling assumptions*
  - ▷ *Longitudinal profiles shape*
    - * non-linear subject-specific trajectories
  - ▷ *Functional form*
    - * how to link the hazard of the event with the longitudinal outcome

Joint Models Framework - Basic Idea

- ▷ Use a model to describe the subject-specific longitudinal trajectories
- ▷ Use these trajectories in a hazard model for the event
- ▷ Random effects explain the association

2 Joint Models (cont'd)



2 Joint Models (cont'd)

More formally

$$
\begin{cases}
h_i(t \mid \mathcal{H}_i(t, \boldsymbol{b}_i)) = h_0(t) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + f(\alpha, \mathcal{H}_i(t, \boldsymbol{b}_i))\}, \\
\qquad\qquad\qquad\qquad \mathcal{H}_i(t, \boldsymbol{b}_i) = \{\eta_i(s, \boldsymbol{b}_i); 0 \leq s \leq t\} \\
\\
y_i(t) = \eta_i(t, \boldsymbol{b}_i) + \varepsilon_i(t) \\
\qquad\quad = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \\
\\
\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})
\end{cases}
$$

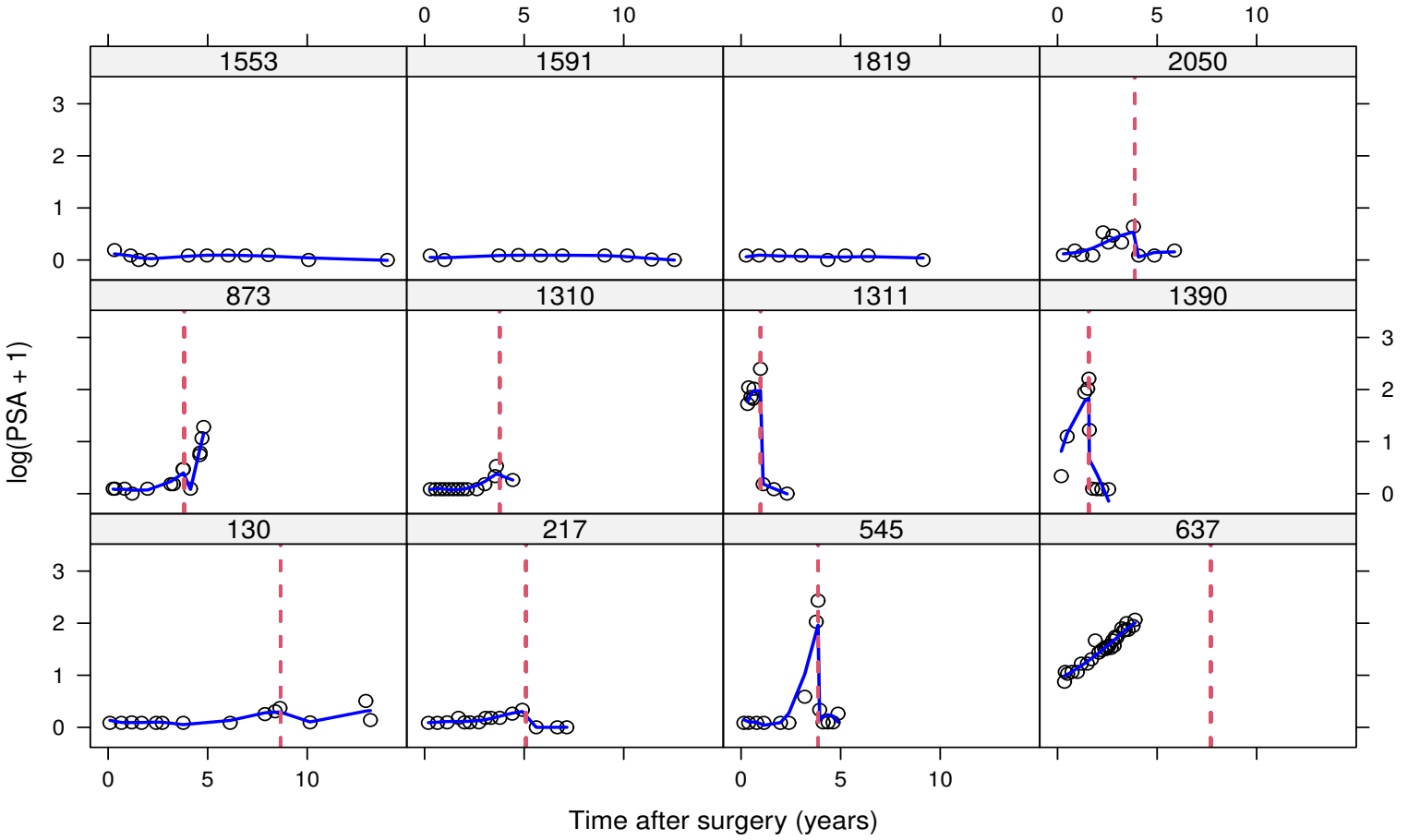
2 Joint Models (cont'd)

- In the context of dynamic predictions,
  - ▷ previous research has shown that predictive accuracy is compromised
  - ▷ when the model does not *adequately* capture the subject-specific trajectories shape

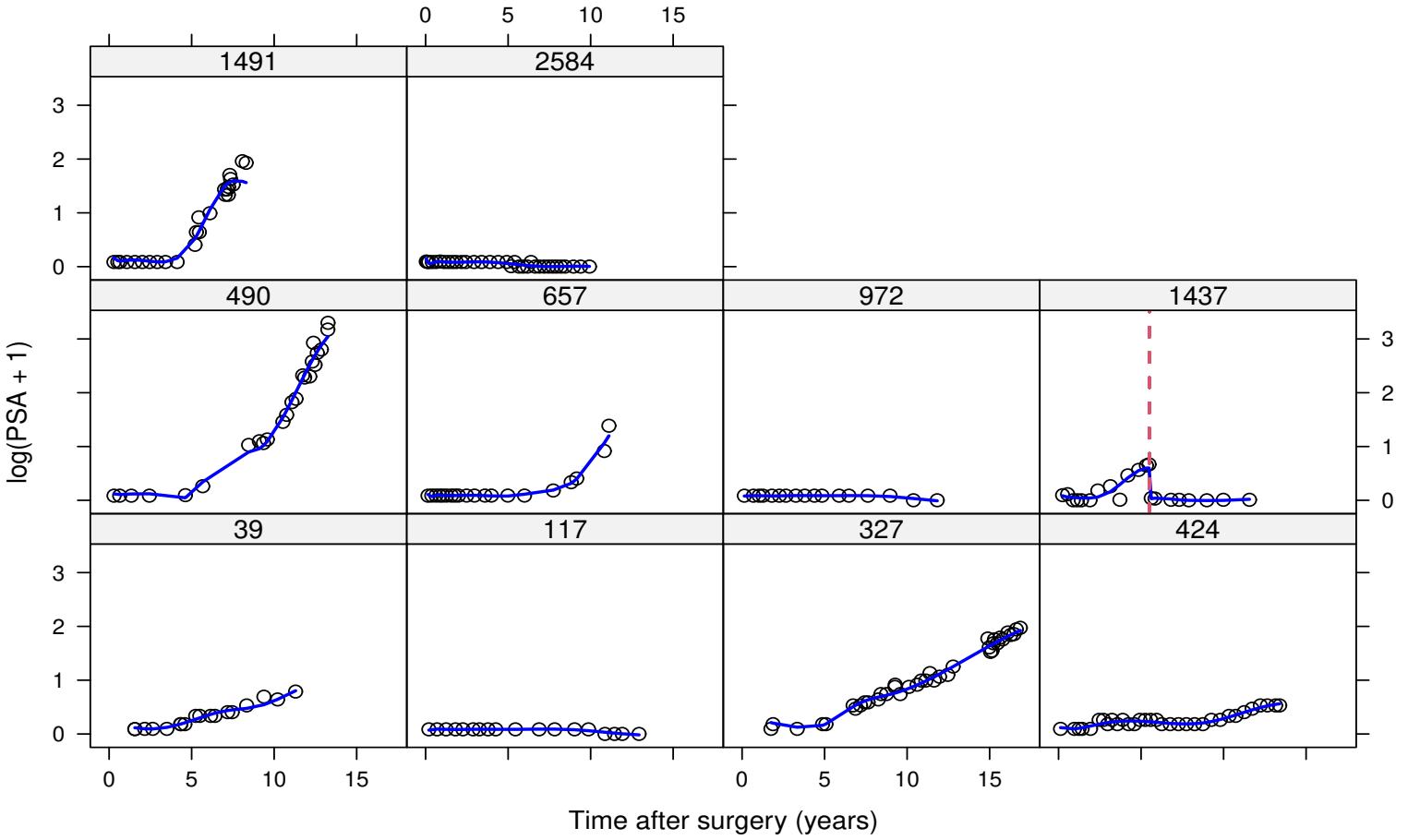
Advice

- ▷ use flexible models, e.g., splines in both fixed- and random-effects parts
- ▷ *increased computational burden*

2 Joint Models (cont'd)



2 Joint Models (cont'd)



3 Super Learning

- The selected functional form and time effect for the longitudinal outcome can influence the derived predictions
  - ▷ especially for the survival outcome

How to select between the different functional forms and trajectory shapes?

3 Super Learning (cont'd)

- The standard answer is to employ information criteria, e.g., DIC, WAIC, ...

- However, the longitudinal information dominates the joint likelihood
  $\Rightarrow$ will not be sensitive enough wrt predicting survival probabilities

- In addition, will *a single model* be the most appropriate
  - ▷ for all future patients?
  - ▷ for all follow-up times?

3 Super Learning (cont'd)

Solution: *Super Learning*

- *Consider multiple plausible models with different*
  - longitudinal outcomes
  - assumptions for the longitudinal profiles
  - functional forms
  - baseline covariates, interaction terms
  - ...

- *Obtain the desired predictions from these models*

- *Combine predictions using weights*
  - *choice of the weights is important*

3 Super Learning (cont'd)

Solution: *Super Learning*

- ▷ select weights to optimize prediction metric *of your choice*
- ▷ account for over-fitting using cross-validation

3 Super Learning (cont'd)

How it works:

- Assume we have a library of L *base-learners* (models) $\mathcal{L} = \{M_1, \ldots, M_L\}$
- Specify time t to optimize the dynamic predictions, and a medically-relevant time window Δt
- Split $\mathcal{D}_n$ in V-folds
- For $v \in \{1, \ldots, V\}$, fit the learners in library $\mathcal{L}$ in $\mathcal{D}_n^{(-v)}$

3 Super Learning (cont'd)

How it works:

▷ For the subjects in $\mathcal{D}_n^{(v)}$, not used when fitting the learner, calculate the predictions

$$\hat{\pi}_i^{(v)}(t + \Delta t \mid t, M_l) = \Pr\{T_i^* < t + \Delta t \mid T_i^* > t, \mathcal{H}_i(t), M_l, \mathcal{D}_n^{(-v)}\}$$

do this for all $v = 1, \ldots, V$ to get the *cross-validated predictions*

3 Super Learning (cont'd)

How it works:

▷ We define the ensemble of *cross-validated predictions*

$$\hat{\bar{\pi}}_i^v(t+\Delta t \mid t) = \sum_{l=1}^{L} \varpi_l(t)\hat{\pi}_i^{(v)}(t+\Delta t \mid t, M_l), \quad v = 1, \ldots, V$$

* the weights depend on $t \Rightarrow$ *different weights at different follow-up times*

3 Super Learning (cont'd)

How it works:

- Select $\varpi_l(t)$ to optimize your *meta-learner* (predictive accuracy metric), e.g.,
  - Brier Score (*Proper scoring rule*)
  - Expected Predictive Cross-Entropy (*Proper scoring rule*)
  - AUC (*Not a proper scoring rule*)
  - ...

- Under the constraints
  - $\widehat{\varpi}_l(t) > 0$ for all $l = 1, \ldots, L$
  - $\sum_{l=1}^{L} \widehat{\varpi}_l(t) = 1$

3 Super Learning (cont'd)

How it works:

- ▷ We obtain $\widehat{\varpi}_l(t)$ using a general-purpose optimization algorithm (e.g., `optim()` in R)
  - * we can transform to an unconstraint problem using the logistic transformation

4 Choice of the Meta-Learner

We focus on two meta-learners

▷ *Brier Score*

$$\text{BS}(t+\Delta t, t) = E\left[\left\{\mathbb{I}(T_i^* \leq t+\Delta t) - \pi_i(t+\Delta t \mid t)\right\}^2 \,\middle|\, T_i^* > t\right]$$

▷ *Expected Predictive Cross-Entropy*

$$\text{EPCE}(t+\Delta t, t) = E\left\{-\log\left[p\left\{T_i^* \mid t < T_i^* \leq t+\Delta t, \mathcal{Y}_i(t)\right\}\right]\right\}$$

5 UM Data Analysis

A library $\mathcal{L}$ with twelve joint models

- PSA models
  - ▷ M_{l1}: *linear* subject-specific time trends that change after salvage
  - ▷ M_{l2}: the same as M_{l1} + covariates
  - ▷ M_{l3}: *nonlinear* subject-specific time trends that change after salvage
  - ▷ M_{l4}: the same as M_{l3} + covariates
- Baseline covariates: age at surgery, Charlson's index, Gleason score, and baseline PSA

5 UM Data Analysis (cont'd)

A library $\mathcal{L}$ with twelve joint models

- Metastasis models
  - ▷ M_{s1}: value of $\log(\text{PSA} + 1)$
  - ▷ M_{s2}: velocity of $\log(\text{PSA} + 1)$
  - ▷ M_{s3}: average $\log(\text{PSA} + 1)$
- Time varying salvage therapy
- Baseline covariates: the same as in the PSA models

5 UM Data Analysis (cont'd)

- We evaluated predictive accuracy in two time intervals
  - ▷ $(4, 7]$: 2514 patients at risk; 28 metastasis
  - ▷ $(6, 9]$: 1914 patients at risk; 16 metastasis

- Metrics
  - ▷ Integrated Brier Score
  - ▷ Expected Predictive Cross-Entropy

5 UM Data Analysis (cont'd)

Results Summary

- Integrated Brier Score
  - no substantive differences between the models
  - for both time intervals

- Expected Predictive Cross-Entropy
  - more sensitive in quantifying differences between models
  - stacking resulted in smaller cross-entropy values
  - the models with nonlinear time trends and the velocity and average PSA functional form dominated the weights

6 Software & Extensions

- Available in **JMbayes2**
  - ▷ cross-validated fitting of models
  - ▷ combination of dynamic predictions
    https://drizopoulos.github.io/JMbayes2/articles/Super_Learning.html

- Causal inference
  - ▷ established theory for combining super learners with *Targeted Maximum Likelihood*

Thank for your attention!

https://www.drizopoulos.com/

6 Super Learning (cont'd)

- Brier Score with IPCW

$$\widehat{\mathsf{BS}}_{IPCW}(t+\Delta t,t)=\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i(t+\Delta t,t)\Big\{\mathbb{I}(T_i\leq t+\Delta t)-\hat{\hat{\pi}}_i^v(t+\Delta t\mid t)\Big\}^2$$

where

$$\widehat{W}_i(t+\Delta t,t)=\frac{\mathbb{I}(t<T_i\leq t+\Delta t)\delta_i}{\hat{G}(T_i\mid t)}+\frac{\mathbb{I}(T_i>t+\Delta t)}{\hat{G}(t+\Delta t\mid t)},$$

with $\hat{G}(\cdot)$ denoting Kaplan-Meier estimate of the censoring distribution $\Pr(C_i>t)$

6 Super Learning (cont'd)

- Brier Score with model-weights

$$\widehat{\mathsf{BS}}_{model}(t+\Delta t,t)=\frac{1}{n_t}\sum_{i:T_i>t}\delta_i\mathbb{I}(T_i\leq t+\Delta t)\Big\{1-\hat{\tilde{\pi}}_i^v(t+\Delta t\mid t)\Big\}^2$$
$$+\,\mathbb{I}(T_i>t+\Delta t)\Big\{\hat{\tilde{\pi}}_i^v(t+\Delta t\mid t)\Big\}^2$$
$$+(1-\delta_i)\mathbb{I}(T_i\leq t+\Delta t)\Big[\hat{\tilde{\pi}}_i^v(t+\Delta t\mid T_i)\Big\{1-\hat{\tilde{\pi}}_i^v(t+\Delta t\mid t)\Big\}^2$$
$$+\Big\{1-\hat{\tilde{\pi}}_i^v(t+\Delta t\mid T_i)\Big\}\Big\{\hat{\tilde{\pi}}_i^v(t+\Delta t\mid t)\Big\}^2\Big]$$

6 Super Learning (cont'd)

- IPCW
  - ▷ *Advantage:* it provides unbiased estimates even when the model is misspecified
  - ▷ *Disadvantage:* it requires that the model for the weights is correct
    * in settings where joint models are used, challenging because censoring may depend on the longitudinal outcomes in a complex manner

6 Super Learning (cont'd)

- Model-based Weights
  - ▷ *Advantage:* it allows censoring to depend on the longitudinal history (in any possible manner)
  - ▷ *Disadvantage:* it requires that the model is well calibrated