Optimizing Dynamic Predictions from Joint Models Using Super Learning

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`

1 Background & Motivation

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

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 dynamically predict the metastasis risk

1 Background & Motivation (cont'd)

- Two main frameworks to obtain such predictions

  ▷ *Landmarking*

  * a series of Cox models at different landmark times

  * biomarker last value as a baseline covariate or a mixed model

  * 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

2 Joint Models

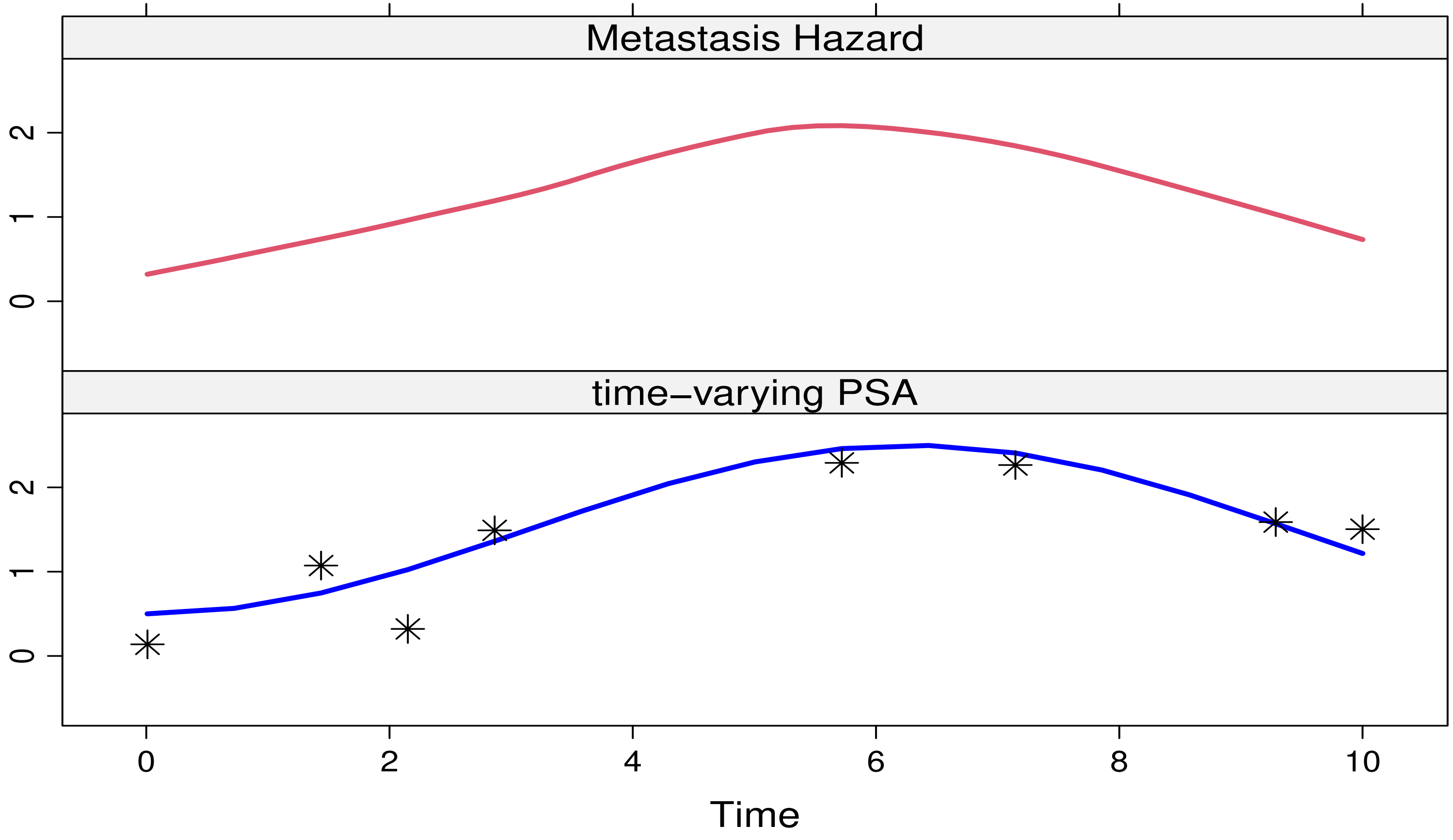
Joint Models Framework - Basic Idea

▷ 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

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

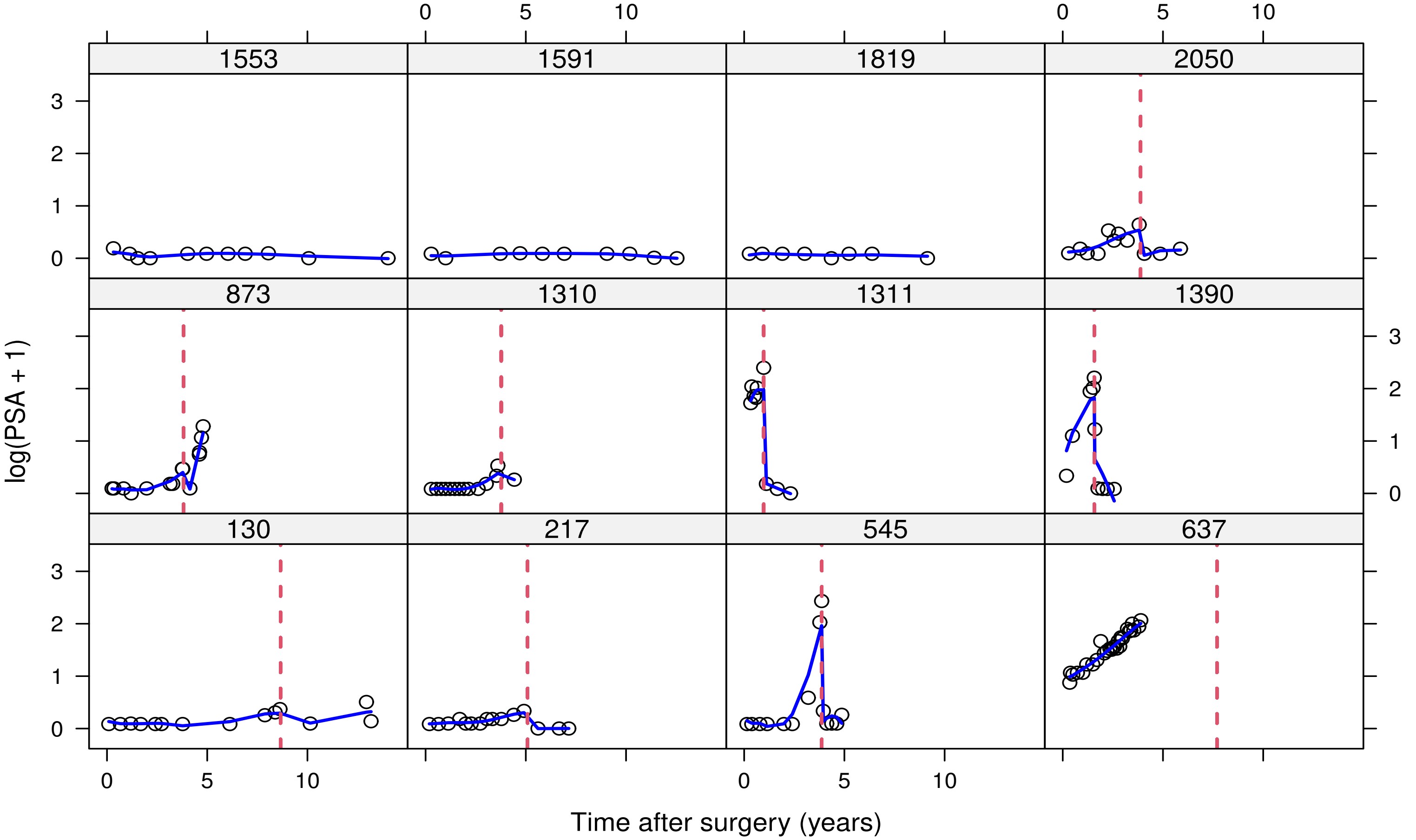
2 Joint Models (cont'd)

- 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

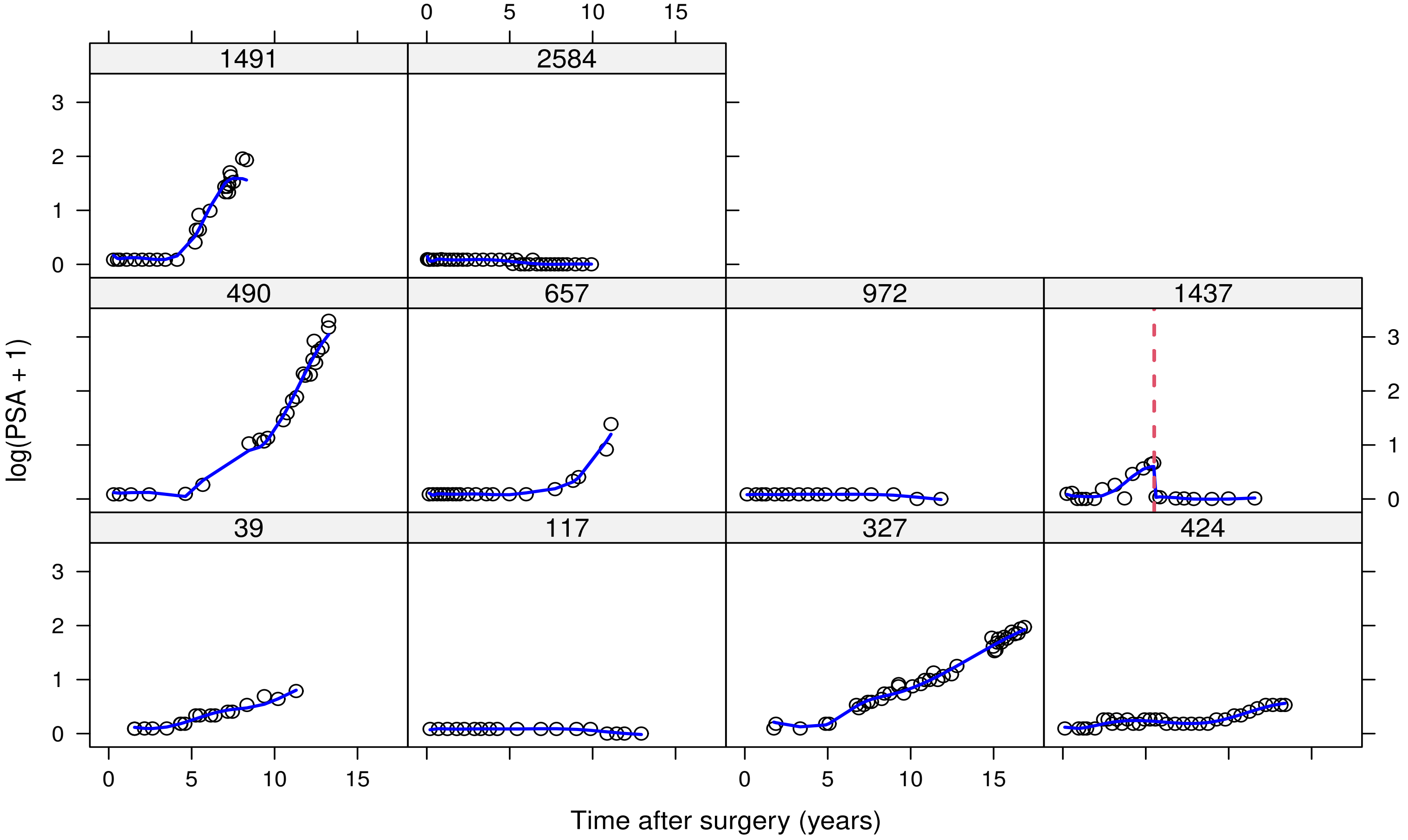
Advice

- ▷ 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 Functional Forms

There are different ways to link the longitudinal trajectories to the risk of an event

▷ Some standard options are . . .

3 Functional Forms (cont'd)

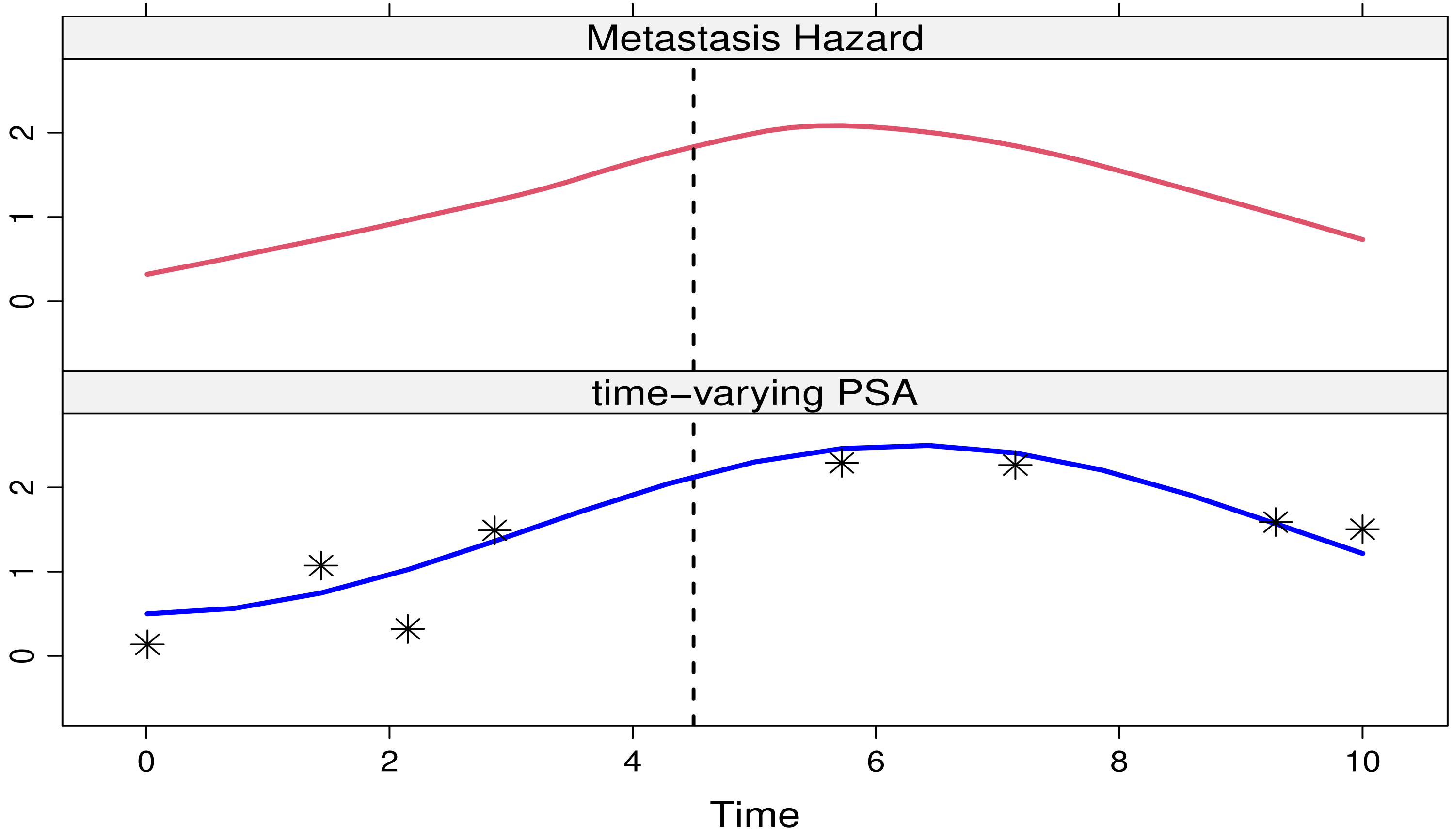
Value: The hazard of metastasis at t is associated with the level of PSA at t:

$$h_i(t \mid \mathcal{H}_i(t, \boldsymbol{b}_i)) \; = \; h_0(t) \exp\bigl\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha \eta_i(t, \boldsymbol{b}_i)\bigr\}$$

where

$$\eta_i(t, \boldsymbol{b}_i) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i$$

3 Functional Forms (cont'd)



3 Functional Forms (cont'd)

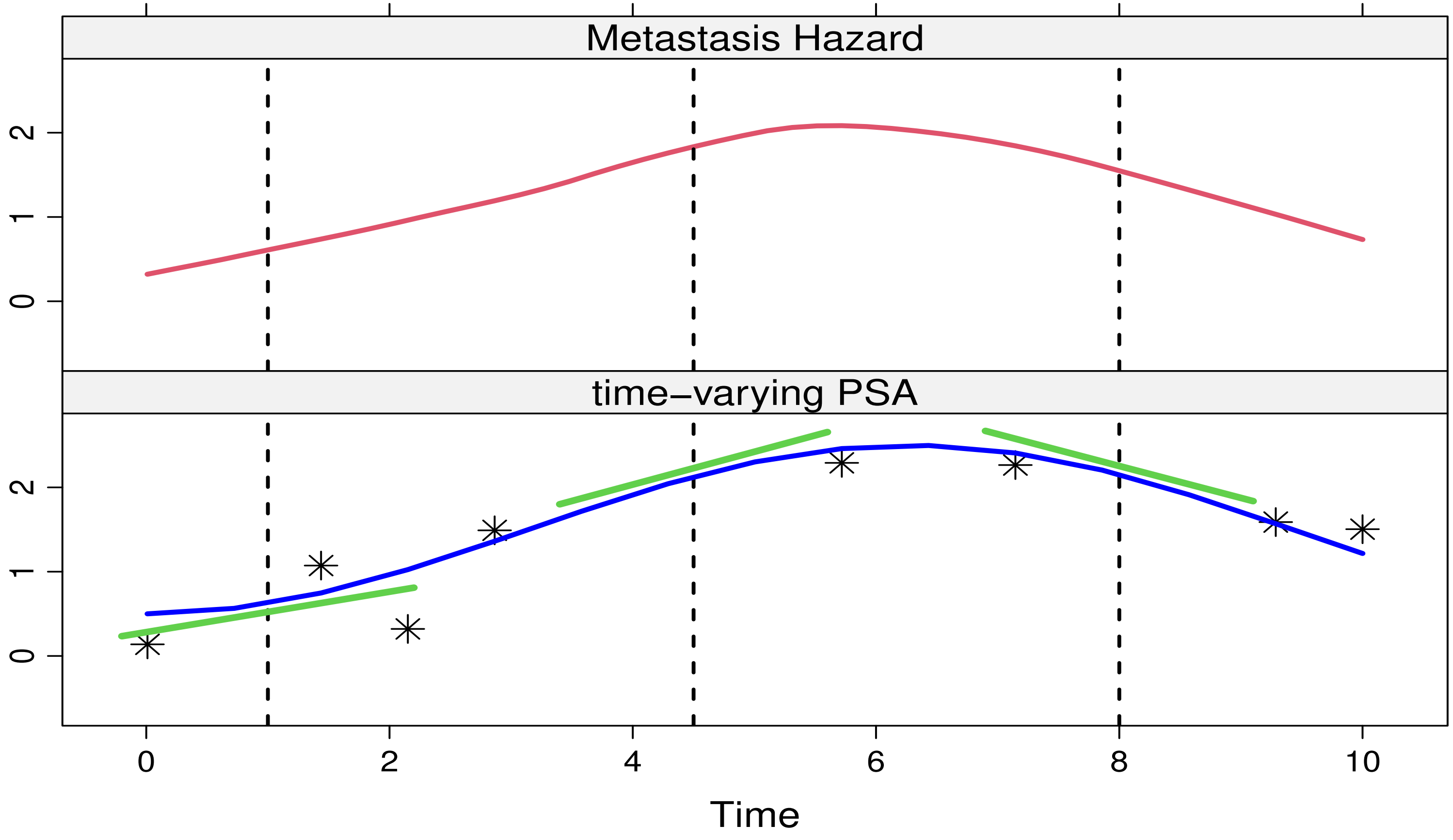
Velocity: The hazard of metastasis at t is associated with the slope of the PSA trajectory at t:

$$h_i(t \mid \mathcal{H}_i(t, \boldsymbol{b}_i)) \; = \; h_0(t) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha \eta_i'(t, \boldsymbol{b}_i)\},$$

where

$$\eta_i'(t, \boldsymbol{b}_i) = \frac{\mathsf{d}}{\mathsf{d}t}\{\boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i\}$$

3 Functional Forms (cont'd)



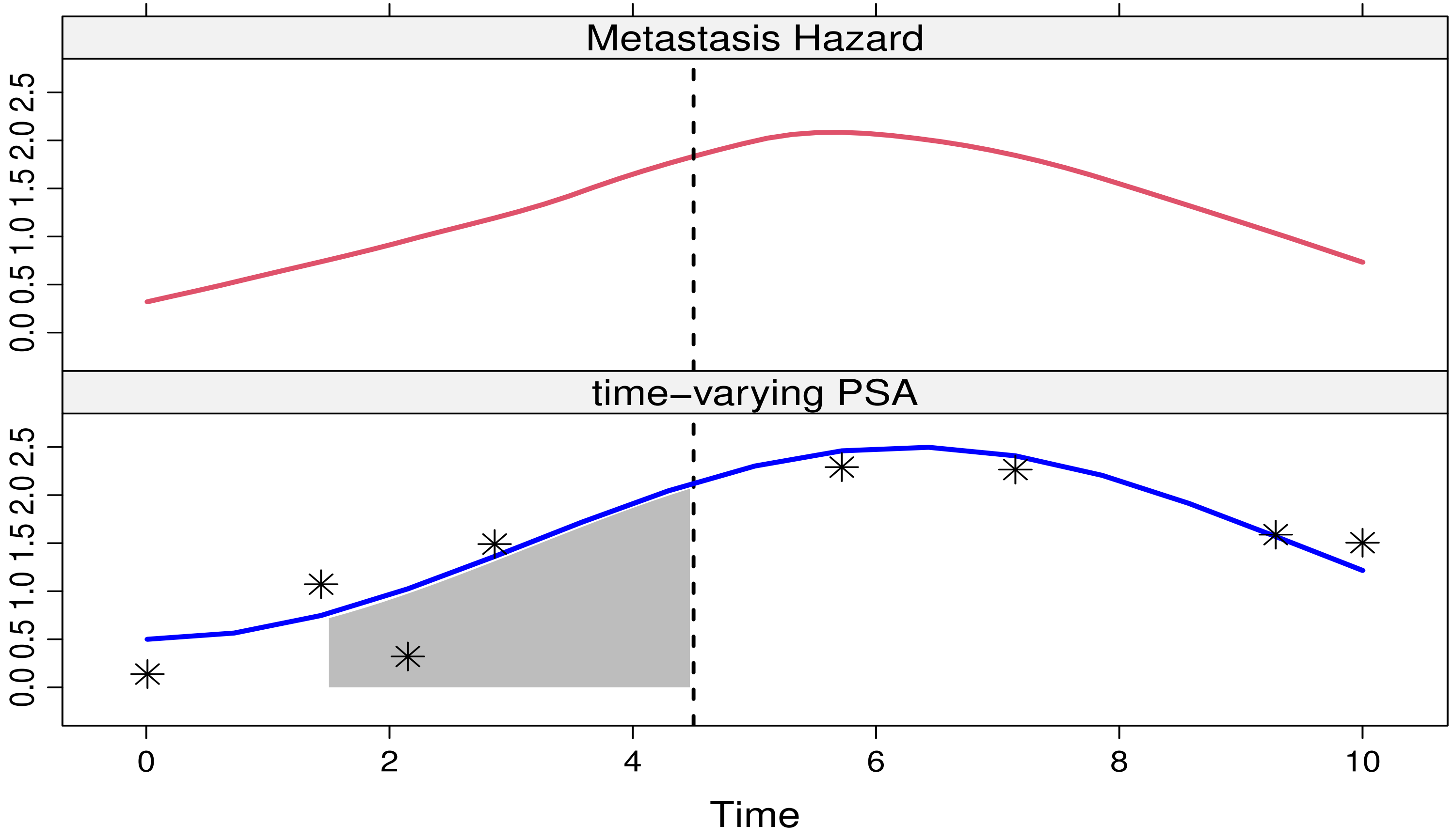
3 Functional Forms (cont'd)

Average Effects: The hazard of metastasis at t is associated with the average PSA in the interval $(t - \Delta t, t)$:

$$h_i(t \mid \mathcal{H}_i(t, \boldsymbol{b}_i)) \; = \; h_0(t) \exp\Big\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha \frac{1}{\Delta t} \int_{t-\Delta t}^{t} \eta_i(s, \boldsymbol{b}_i) \, \mathrm{d}s \Big\}$$

We account for the observation period

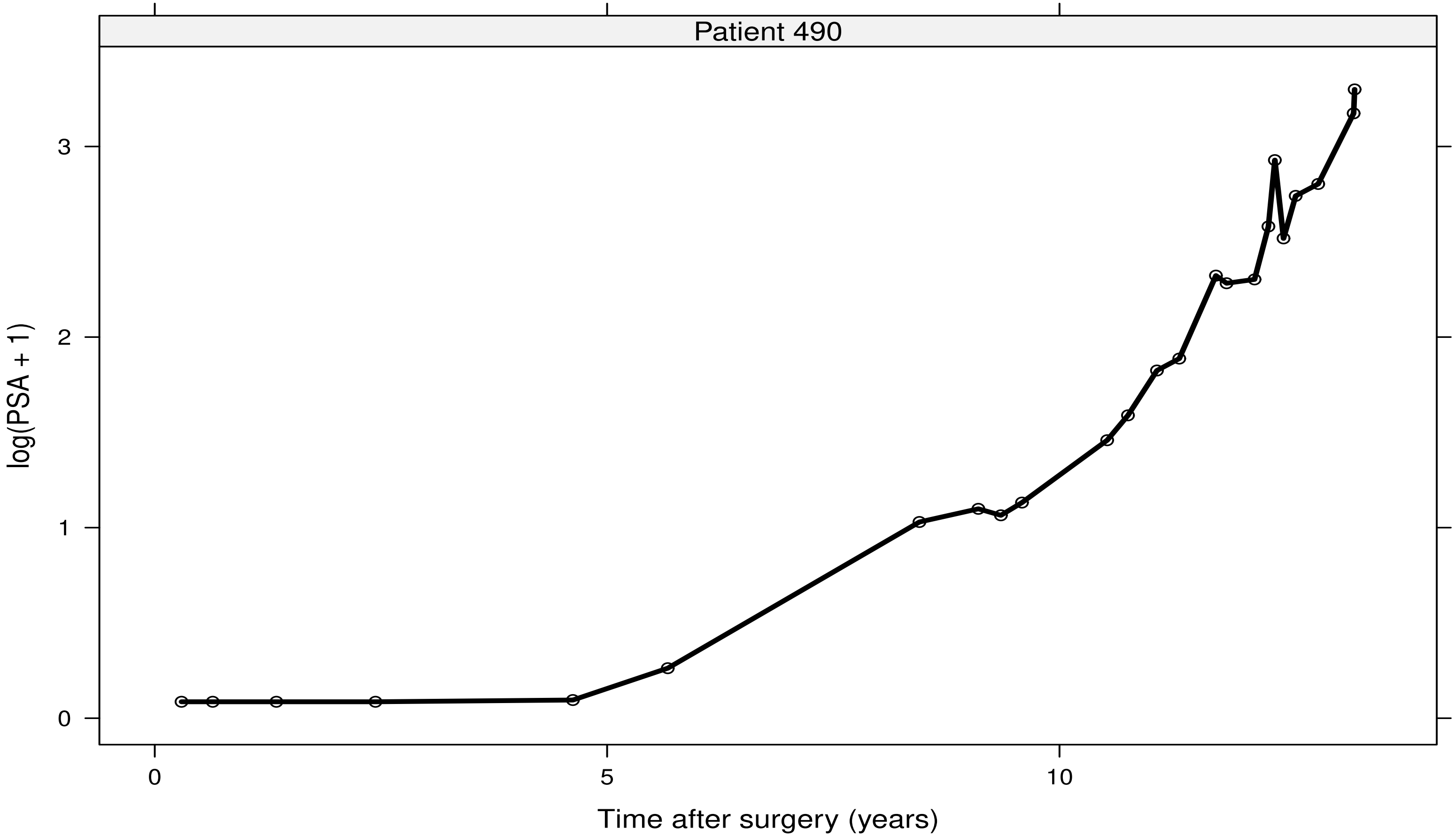
3 Functional Forms (cont'd)



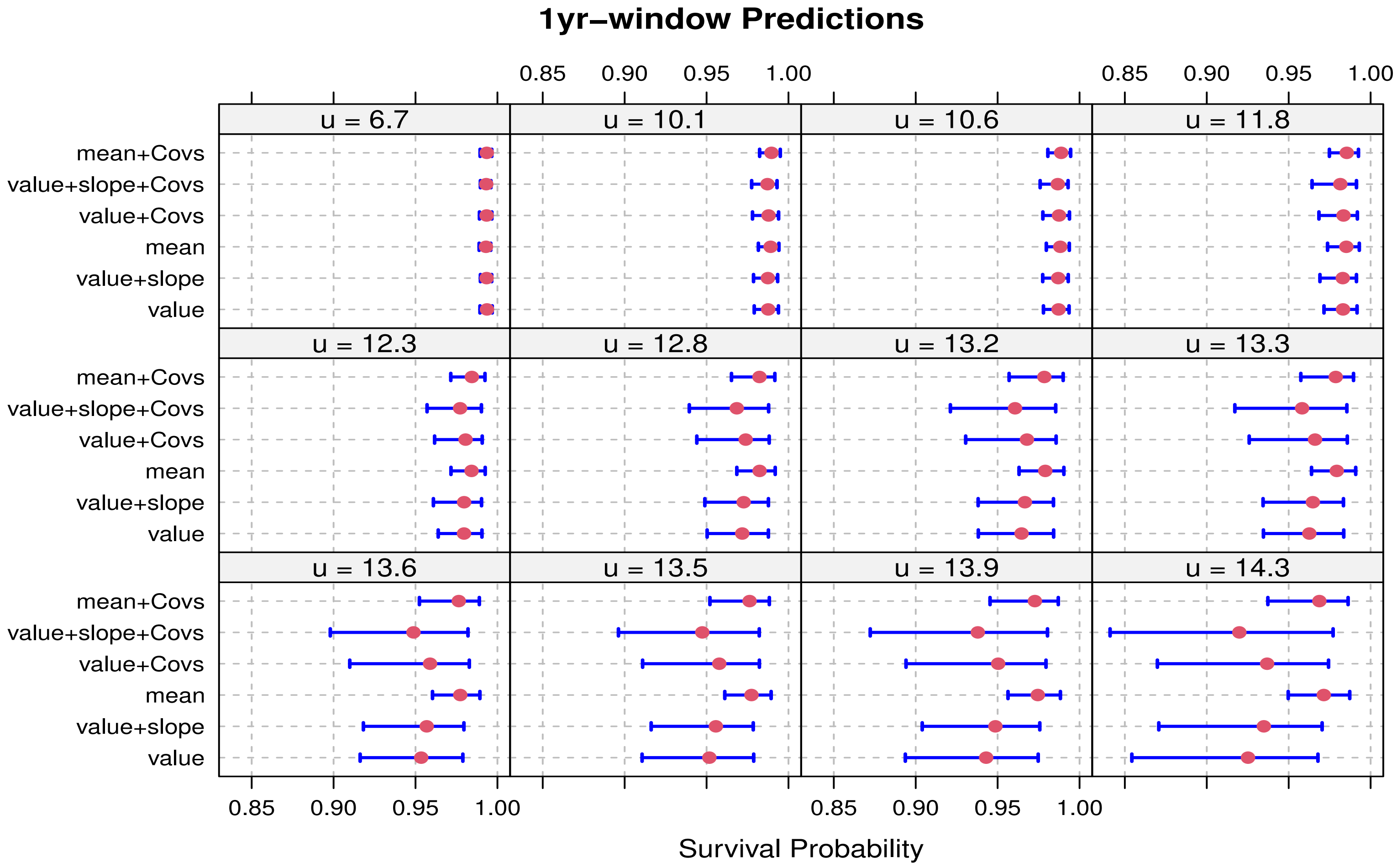
3 Functional Forms (cont'd)

How significant is the choice of the functional form for dynamic predictions?

3 Functional Forms (cont'd)



3 Functional Forms (cont'd)



4 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?

4 Super Learning (cont'd)

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

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

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

4 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*

- * *how to select the weights?*

4 Super Learning (cont'd)

Solution: *Super Learning*

▷ select weights to optimize prediction metric *of your choice*

▷ account for over-fitting using cross-validation

4 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 the landmark time t, and a relevant future time u, $u > t$

▷ Split $\mathcal{D}_n$ in V-folds

▷ For $v \in \{1, \ldots, V\}$, train the learners in library $\mathcal{L}$ using $\mathcal{D}_n^{(-v)}$

4 Super Learning (cont'd)

How it works:

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

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

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

4 Super Learning (cont'd)

How it works:

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

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

* the weights depend on t ⇒ *different weights at different follow-up times*

4 Super Learning (cont'd)

How it works:

- ▷ Select $\varpi_l(t)$ to optimize your *meta-learner* (predictive accuracy metric), e.g.,
  - * Brier Score
  - * Expected Predictive Cross-Entropy
  - * AUC
  - * ...

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

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

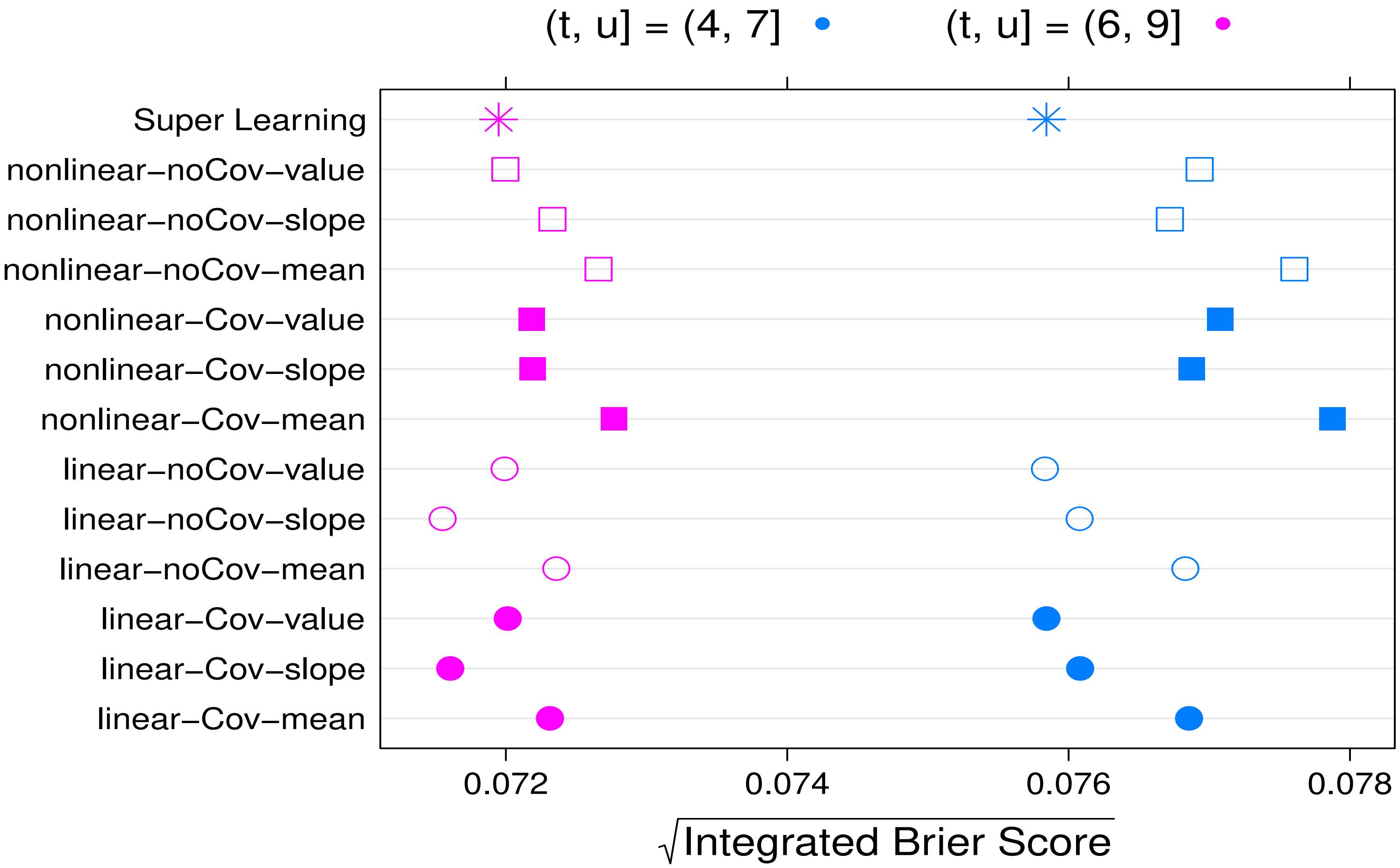
- Baseline covariates: the same as in the PSA models

5 UM Data Analysis (cont'd)

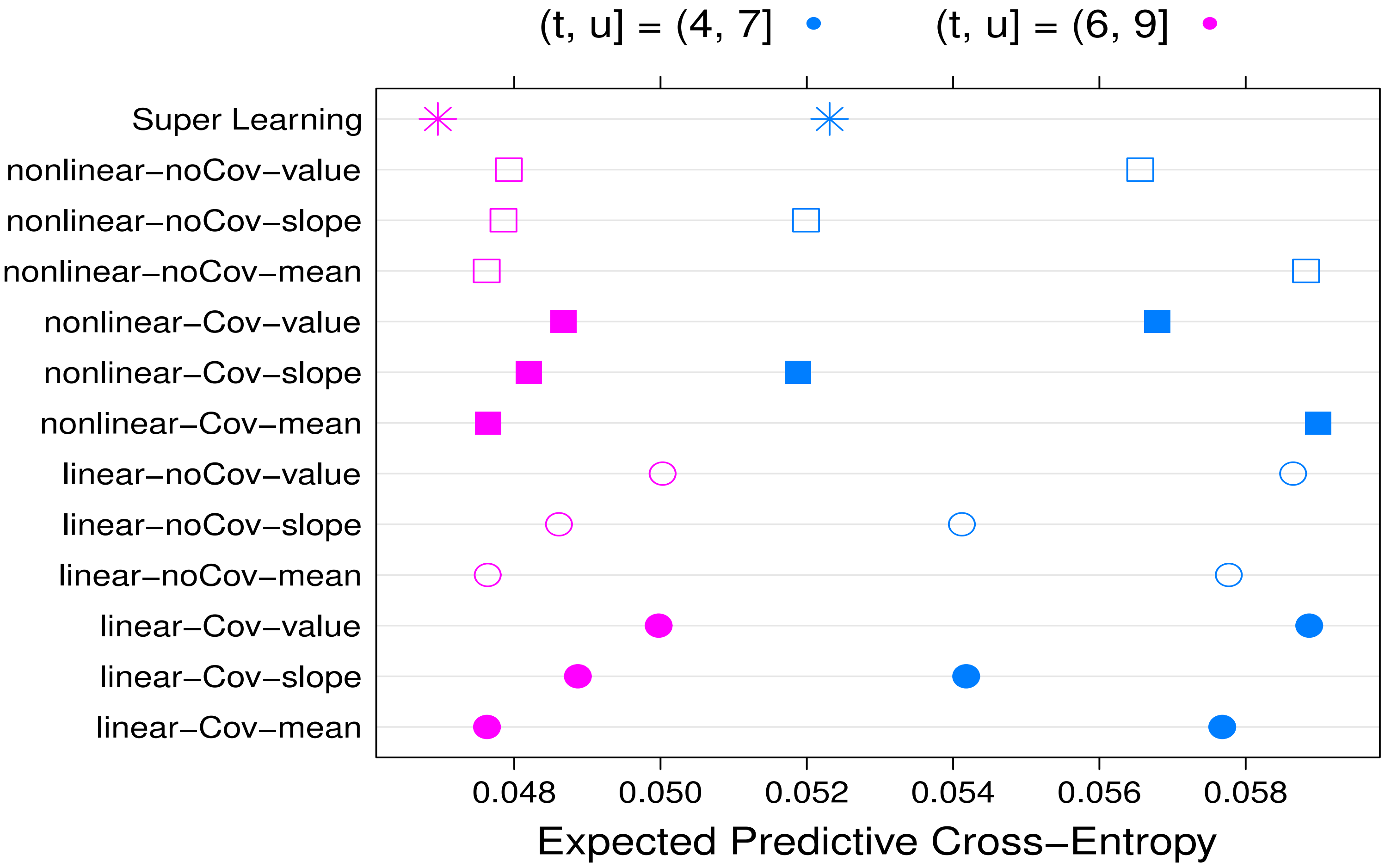
- We evaluated predictive accuracy in two time intervals
  - ▷ $(4, 7]$: 2514 patients at risk; 28 metastasis
  - ▷ $(6, 9]$: 1914 patients at risk; 16 metastasis

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

5 UM Data Analysis (cont'd)



5 UM Data Analysis (cont'd)



5 UM Data Analysis (cont'd)

Observations (also from the simulation study)

▷ ensemble Super Learning (eSL) often, *but not always*, outperforms the individual models

▷ In some datasets and intervals $(t, u]$, the discrete Super Learner (dSL) beats the eSL

Recommendation

Regard eSL as an extra member of the library $\mathcal{L}$ and use CV to select the optimal strategy

6 Software

- Available in **JMbayes2**
  - ▷ cross-validated fitting of models
  - ▷ combination of dynamic predictions

    https://drizopoulos.github.io/JMbayes2/articles/Super_Learning.html

Thank for your attention!

https://www.drizopoulos.com/

7 Choice of the Meta-Learner

We focus on two meta-learners

▷ *Integrated Brier Score*

$$\text{IBS}(t+\Delta t, t) = \frac{1}{\Delta t}\int_t^{t+\Delta t} E\left[\left\{\mathbb{I}(T_i^* \leq s) - \pi_i(s \mid t)\right\}^2 \,\Big|\, T_i^* > t\right] \mathrm{d}s$$

▷ *Expected Predictive Cross-Entropy*

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

7 Choice of the Meta-Learner (cont'd)

- For the estimation of the Brier score, we need to account for censoring in $[t, t + \Delta t)$
  - * inverse probability of censoring weighting
  - * model-based weights

7 Choice of the Meta-Learner (cont'd)

- Brier Score with IPCW

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

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 $\Pr(C_i > t)$

7 Choice of the Meta-Learner (cont'd)

- Brier Score with model-weights

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

$$+ \, \mathbb{I}(T_i > t + \Delta t) \Big\{ \hat{\tilde{\pi}}_i^v (t + \Delta t \mid t) \Big\}^2$$

$$+ (1 - \delta_i) \mathbb{I}(T_i \leq t + \Delta t) \Big[\hat{\tilde{\pi}}_i^v (t + \Delta t \mid T_i) \Big\{ 1 - \hat{\tilde{\pi}}_i^v (t + \Delta t \mid t) \Big\}^2$$

$$+ \, \Big\{ 1 - \hat{\tilde{\pi}}_i^v (t + \Delta t \mid T_i) \Big\} \Big\{ \hat{\tilde{\pi}}_i^v (t + \Delta t \mid t) \Big\}^2 \Big]$$

7 Choice of the Meta-Learner (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
    - * challenging because censoring may depend on the longitudinal outcomes in a complex manner
    - * sensitive to (unobserved) instrument by confounder interactions

7 Choice of the Meta-Learner (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-specified

7 Choice of the Meta-Learner (cont'd)

- An estimate of EPCE$(t + \Delta t, t)$ that accounts for censoring

$$\widehat{\text{EPCE}}(t + \Delta t, t) = \frac{1}{n_t} \sum_{i: T_i > t} - \log \Bigl[p \bigl\{ \tilde{T}_i, \tilde{\delta}_i \mid T_i > t, \mathcal{Y}_i(t), \mathcal{D}_n \bigr\} \Bigr]$$

with

▷ $\tilde{T}_i = \min(T_i, t + \Delta t)$

▷ $\tilde{\delta}_i = \delta_i \mathbb{I}(t < T_i \leq t + \Delta t)$

- Features

▷ it allows censoring to depend on the longitudinal history

▷ *problem:* it is not written as a function of the predictions

7 Choice of the Meta-Learner (cont'd)

- The conditional predictive log-likelihood

$$\log\Big[p\big\{\tilde T_i, \tilde\delta_i \mid T_i > t, \mathcal{Y}_i(t), \mathcal{D}_n\big\}\Big] =$$

$$\tilde\delta_i \log[h_i\{\tilde T_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}] + \log \frac{\Pr\{T_i^* > \tilde T_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}{\Pr\{T_i^* > t \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}$$

  ▷ the second term is $\log\{\pi_i(\tilde T_i \mid t)\}$

  ▷ for the first term, we write the hazard function as

$$h_i\{\tilde T_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\} = \frac{p(\tilde T_i)}{S(\tilde T_i)} = -\frac{\frac{\mathsf{d}}{\mathsf{d}t} \Pr\{T_i^* > t \mid \mathcal{Y}_i(t), \mathcal{D}_n\}\Big|_{t=\tilde T_i}}{\Pr\{T_i^* > \tilde T_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}$$

7 Choice of the Meta-Learner (cont'd)

- We approximate the derivative with a forward difference and we get

$$\widehat{\mathsf{EPCE}}(t + \Delta t, t) =$$

$$-\frac{1}{n_t} \sum_{i:T_i > t} \tilde{\delta}_i \big[\log\{1 - \hat{\tilde{\pi}}_i^v(\tilde{T}_i + \epsilon \mid \tilde{T}_i)\} - \log(\epsilon) \big] + \log\{\hat{\tilde{\pi}}_i^v(\tilde{T}_i \mid t)\}$$

that can be used to optimize $\varpi_l(t)$

7 UM Data Analysis (cont'd)

	$(t, t + \Delta t] = (4, 7]$		$(t, t + \Delta t] = (6, 9]$	
	IBS	weights	IBS	weights
SL	0.07584		0.07195	
linear-noCov-value	0.07583	0.00000	0.07199	0.08333
linear-noCov-slope	0.07608	0.00000	0.07155	0.08340
linear-noCov-mean	0.07683	0.00000	0.07236	0.08332
linear-Cov-value	0.07584	1.00000	0.07201	0.08335
linear-Cov-slope	0.07608	0.00000	0.07160	0.08339
linear-Cov-mean	0.07686	0.00000	0.07231	0.08332
nonlinear-noCov-value	0.07693	0.00000	0.07200	0.08334
nonlinear-noCov-slope	0.07672	0.00000	0.07233	0.08331
nonlinear-noCov-mean	0.07760	0.00000	0.07266	0.08329
nonlinear-Cov-value	0.07708	0.00000	0.07218	0.08332
nonlinear-Cov-slope	0.07687	0.00000	0.07219	0.08333
nonlinear-Cov-mean	0.07788	0.00000	0.07277	0.08328

7 UM Data Analysis (cont'd)

	$(t, t + \Delta t] = (4, 7]$		$(t, t + \Delta t] = (6, 9]$	
	EPCE	weights	EPCE	weights
SL	0.05231		0.04696	
linear-noCov-value	0.05865	0.08325	0.05003	0.00002
linear-noCov-slope	0.05412	0.08320	0.04861	0.00000
linear-noCov-mean	0.05777	0.08260	0.04764	0.39649
linear-Cov-value	0.05887	0.08215	0.04997	0.00000
linear-Cov-slope	0.05418	0.08333	0.04887	0.00000
linear-Cov-mean	0.05768	0.08270	0.04763	0.12793
nonlinear-noCov-value	0.05656	0.08337	0.04793	0.00136
nonlinear-noCov-slope	0.05199	0.08517	0.04785	0.44966
nonlinear-noCov-mean	0.05882	0.08296	0.04762	0.00961
nonlinear-Cov-value	0.05679	0.08315	0.04867	0.00000
nonlinear-Cov-slope	0.05188	0.08526	0.04820	0.01327
nonlinear-Cov-mean	0.05899	0.08288	0.04764	0.00166