# Joint models for longitudinal and survival data, with application to prostate cancer.

*Jeremy M. G. Taylor*
*Department of Biostatistics*
*University of Michigan*

*Joint work with*
*Menggang Yu, Wen Ye, Xihong Lin*

## Outline

- INTRODUCTION
- DESCRIPTION OF JOINT MODELS
- PROSTATE CANCER APPLICATION
- SEMI-PARAMETRIC LONGITUDINAL MODEL
- OPEN RESEARCH ISSUES

- Setting: Clinical trial or Observational study

- Biomarker, longitudinal variable
  - Internal covariate (a measure of disease progression)
  - Ex1. CD4 and Viral load in HIV studies
  - Ex2. PSA in prostate therapy studies
  - Continuous variable in longitudinal model
- Clinically important endpoint
  - Occurence of AIDS, death from AIDS
  - Recurrence of prostate cancer following treatment
  - Censored event time in survival model

- Data
  - $(t_i, \delta_i)$, censored event time
  - $X_i$, time-independent covariates
  - $Y_{ij}$, time-dependent covariate, biomarker

- Both T and Y are response variables
- X could be treatment group or stage of disease

TIME SEQUENCE

1. Intervention or Exposure, X
2. Longitudinal Biomarker, Y
3. Clinical Event, T

PRIOR KNOWLEDGE

- From science/biology and preliminary data
- Expect Y to be affected by X
- Y associated with T

MY APPROACH TO MODELLING LONGITUDINAL DATA

1. There is an underlying multivariate stochastic process that generated the data
2. Goal of modelling is to describe and understand the stochastic process
3. Scientific context and prior similar data may suggest some reasonable assumptions, for example
   - Smoothness, monotonicity
   - Unimodal distributions
   - Exponential growth
   - Transformations
4. Model should be faithful to the time sequence
5. A variety of conclusions and inferences follow from the descriptive model

6. Statistical parsimony

7. Efficiency matters as well as bias

8. Data should fit the model

- General model for $[T, Y \mid X]$
  - Factor as $[Y \mid X][T \mid Y, X]$
- $[Y \mid X]$, longitudinal model
  - random effects
  - measurement error
  - unbalanced time of observations
- $[T \mid Y, X]$, survival model
  - time-dependent Cox model
  - Y not fully observed
  - dependent censoring

POSSIBLE GOALS

1: Survival analysis, parameters of $[T \mid Y, X]$

2: Longitudinal analysis, parameters of $[Y \mid X]$

3: Estimation of marginal survival distribution, $[T]$ or $[T \mid X]$

4: Use Y as a auxiliary variable to help in the estimation of $[T]$ and $[T \mid X]$

5: Use Y as a surrogate endpoint, instead of T, in a clinical trial.

6: Prediction of future longitudinal and event times for individual patients

GENERIC JOINT MODEL

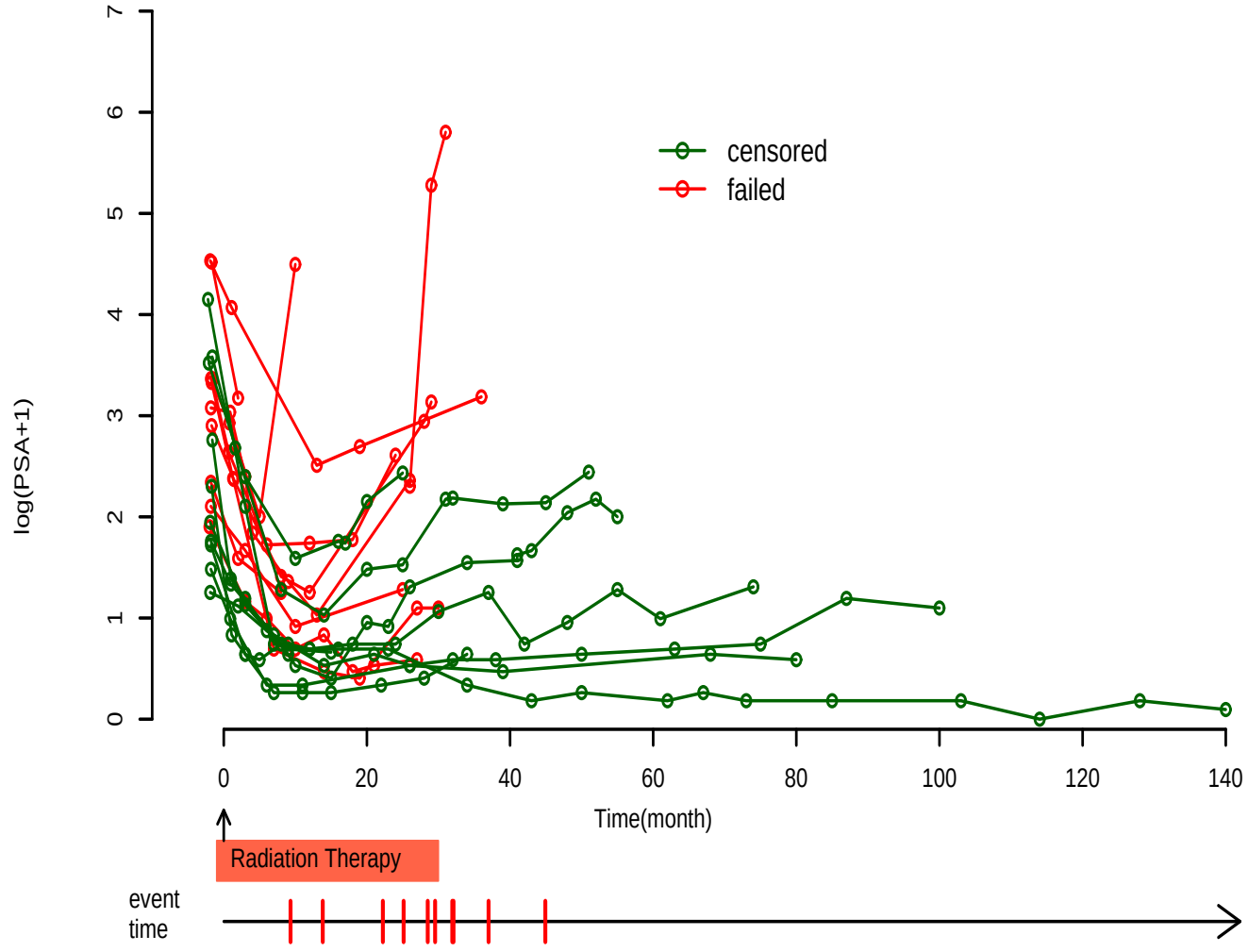
- Longitudinal model (random effects)
  - $Y_i(t_{ij}) = Z_i(t_{ij}) + e_{ij}$
  - $Z_i(t) = X_i\beta + a_i + b_i t$
  - $(a_i, b_i) \sim$ Gaussian
- Hazard model (proportional hazards)
  - $\lambda(t) = \lambda_0(t) exp(\alpha Z_i(t) + \omega X_i)$
- $\lambda_0(t)$ parametric or non-parametric
- Estimation
  - MLE, usually EM algorithm
  - MCMC
  - Computationally intensive, no standard software

**SOME ISSUES**

- Change linear $(a_i + b_i t\ )$ to smooth/stochastic process $(f_i(t))$
- $(a_i, b_i)$ not Gaussian, robustness.
- Hazard depends on more than current value of $Z_i(t)$
  - History of Z $\{Z_i(s), 0 \leq s \leq t\}$
  - Slope of $Z_i(t)$

**Data**

- Patients treated with radiation therapy for prostate cancer (n=921).
- Baseline covariates.
- Longitudinal marker (PSA).
- Censored clinical event times.


censored
failed
log(PSA+1)
0
1
2
3
4
5
6
7
0
20
40
60
80
100
120
140
Time(month)
Radiation Therapy
event
time

## Post-treatment PSA

- measured every 6 months

- a total of 7306 post-treatment PSA values

- median no. of PSA per patient is 7 (range is 1-31)

- the last PSA is measured at around 144 months after treatment

## Endpoints and Censoring

- local recurrence and distant metastasis (n=126)
- censoring: lost to follow-up, end of study, death, start of hormonal therapy (n=795)
- follow-up time: median = 50 months , maximum = 148 months

## Goals

1. Understand the relationship between
   - Baseline variables (Gleason Score, T-stage, bPSA)
   - PSA trajectory
   - Clinical recurrence (Local recurrence, Distant Metastasis)

   in a UNIFIED way.
2. Make individual predictions.
   - for each patient who hasn't been observed to have a clinical event, assign a probability of the patient being cured or the probability of an event in the next 3 years given their baseline variables and history of PSA.

## Idea

- Patient can be 'cured' (D=2) or 'not cured' (D=1). This occurs at the time of radiation therapy.
- What factors influence the probability of cure
- What factors influence the pattern of PSA given cured.
- What factors influence the pattern of PSA given not cured.
- What factors influence the recurrence hazard given not cured.

## Model Specification

Notation

$D_i$ - partially observed latent variable

$D_i = 1$ non cure; $D_i = 2$ cure

$\mathbf{X_i}$ - baseline covariates.

$PSA_i(t)$ - longitudinal PSA data

$\mathbf{R}_i$ - random effects of longitudinal model

1. Incidence (long term clinical cure).

   logistic model

$$\log\left[\frac{P(D_i = 1 \,|\, \mathbf{b}, \mathbf{X}_i)}{1 - P(D_i = 1 \,|\, \mathbf{b}, \mathbf{X}_i)}\right] = b_0 + b_1(T_i = 1)$$

$$+b_2(T_i = 2) + b_3 \log(1 + bPSA_i) + b_4(Gleason)$$

1. Longitudinal. Non-linear random effects models.

$$\log\Big[1 + PSA_i(t)\Big] = \log\Big[1 + r_{i1}e^{-r_{i2}t} + r_{i3}e^{r_{i4}t}\Big] + e_{it}$$

where $r_{i1}, r_{i2}, r_{i3}$ and $r_{i4}$ are the unobserved random effects for

subject $i$ ($r_{i1}, r_{i2}, r_{i3}$ and $r_{i4} > 0$).

Separate models for $D_i = 1$ and $D_i = 2$.

Mean structure of ($r_{i1}, r_{i2}, r_{i3}$ and $r_{i4}$) depend on $X_i$.

1. Latency. Time-dependent proportional hazards for those in the susceptible group.

$$\lambda_i\big(t \,|\, D_i = 1,\, \mathbf{R}_i, \mathbf{X}_i\big) = \lambda_0(t) \exp\big(\psi(\mathbf{R}_i,\, t) + \boldsymbol{\beta}' \mathbf{X}_i\big)$$

where we take

$$\psi(\mathbf{R}_i,\, t) = \alpha_1 \, \log\Big[1 + PSA_i(t)\Big] + \alpha_2 \, SL_i(t)$$

with $SL_i(t) = \partial \log\Big[1 + PSA_i(t)\Big] / \partial t$ is the current slope of $\log\Big[1 + PSA_i(t)\Big]$ at time $t$.

## Computation

Markov chain Monte Carlo based on likelihood and priors

## Predict Recurrence for Censored Patients

- $\boldsymbol{\Omega}$: parameters;
- $\mathbf{Y}$: observed longitudinal data;
- $\mathbf{T}, \boldsymbol{\Delta}$: survival data;
- $t_i$: last contact time for patient $i$;
- $\boldsymbol{\Omega}^{(k)}$: $k^{th}$ draw from the posterior distribution.

For patient $i$, the conditional probability of recurrence within $a$ months $P\big[\, T_i < t_i + a \,|\, \mathbf{Y}, \mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}_i \,\big]$ can be approximated by

$$\frac{1}{m} \sum_{k=1}^{m} P\big[\, T_i > t_i + a \,|\, \boldsymbol{\Omega}^{(k)}, T_i > t_i, \mathbf{X}_i \,\big]$$

Parameters in the failure time model: $\boldsymbol{\beta}, \alpha_1, \alpha_2$

| Parameter | Mean | S.D. | Mean/S.D. |
|---|---|---|---|
| T1 | -1.40 | 0.31 | -4.49 |
| T2 | -0.51 | 0.19 | -2.64 |
| bPSA | 0.11 | 0.10 | 1.11 |
| Gleason | 0.39 | 0.06 | 6.26 |
| log(1+PSA) | -0.02 | 0.02 | -0.73 |
| Slope | 5.32 | 0.42 | 12.78 |

Figure 1: Prediction of future PSA values

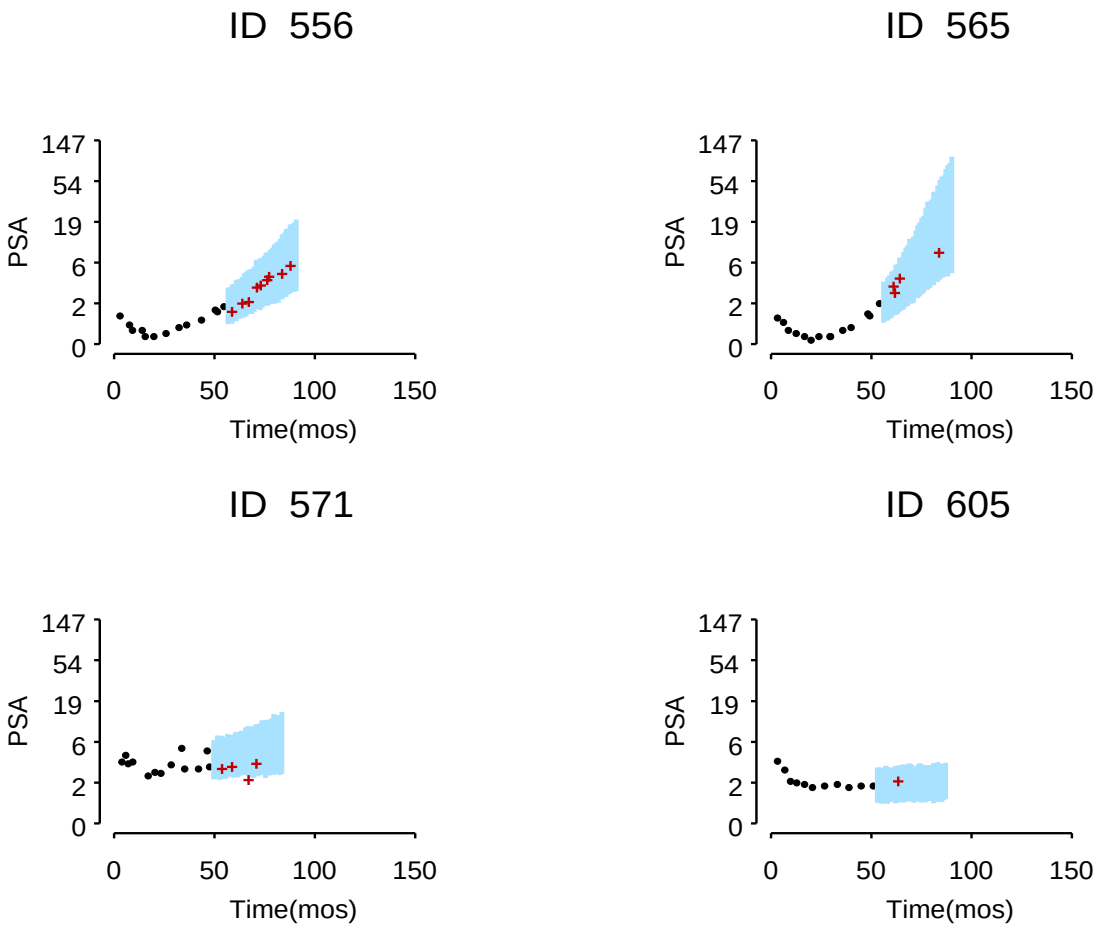
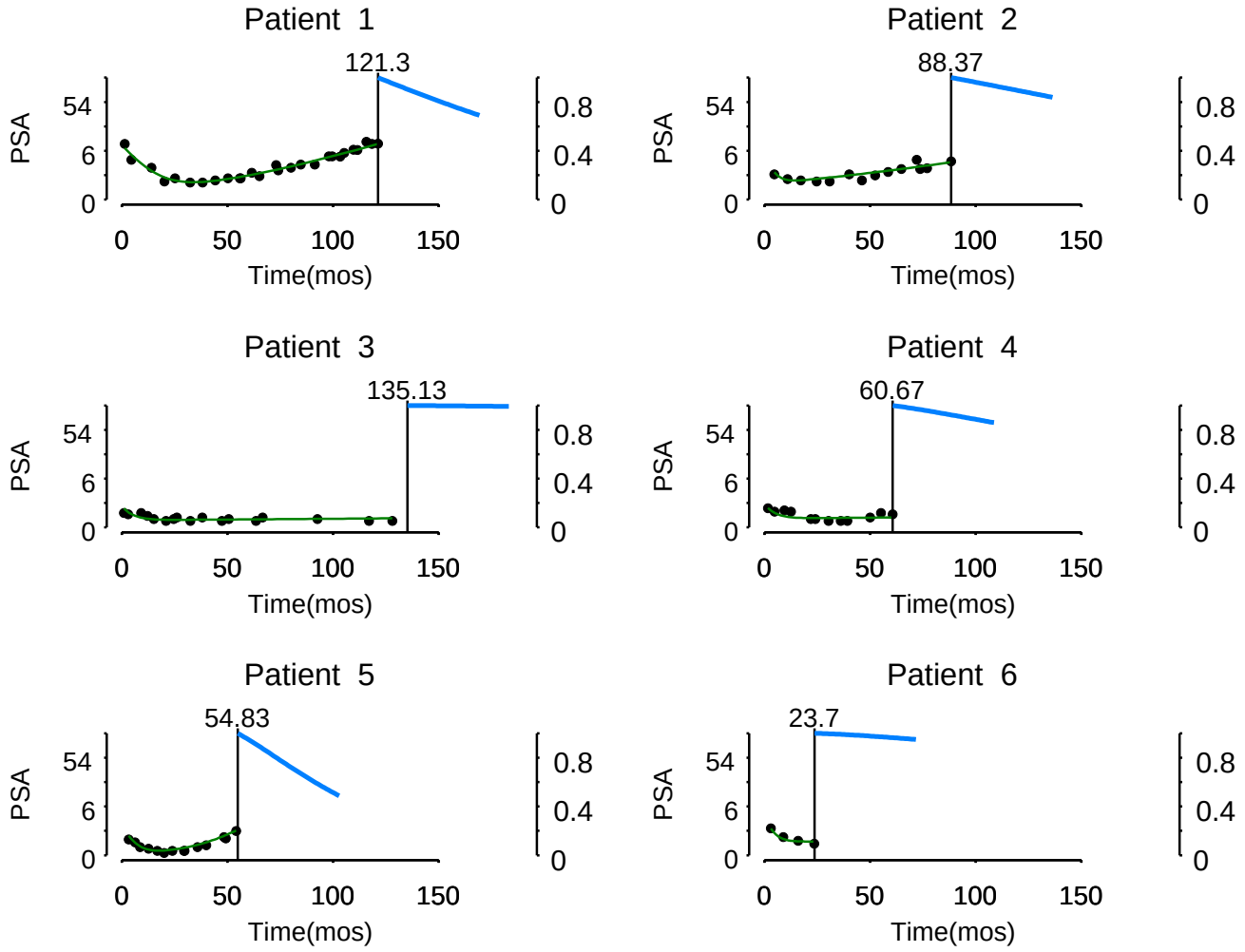


Figure 2: Individual Prediction of clinical recurrence for censored subjects.



## Discussion

- Joint modelling can help understand the relationship between biomarkers and true endpoints
- Reduces bias and gains efficiency compared to separately modelling the two responses
- Models are very parametric
- Checking model fits data is non-trivial.

## Semi-parametric approach

- Smooth Longitudinal model
  - $Y_i(t_{ij}) = Z_i(t_{ij}) + e_{ij}$
  - $Z_i(t) = X\beta + f(t) + W_i(t)$
- Hazard model (proportional hazards)
  - $\lambda(t) = \lambda_0(t)exp(\alpha Z_i(t))$
  - or $\lambda(t) = \lambda_0(t)exp(\alpha_1 Z_i(t) + \alpha_2 SLZ_i(t))$
- $f(t)$ is a smoothing spline, twice differentiable function.
- $W_i(t)$ is integrated Wiener process.
- Can be represented as a mixed model (Wahba, Zhang and Lin)

## Two stage estimation

- Fit longitudinal model, get BLUP estimates of $Z_i(t)$
- Fit hazard model using partial likelihood, with imputed values of $Z_i(t)$

## Three approaches

- LVCF, Naive. Use latest value of Y in Cox partial likelihood.
- ORC, Ordinary regression calibration. Use BLUP estimates based on one fit to all the longitudinal data.
- RRC, Risk set regression calibration. Use BLUP estimate based on past longitudinal data amongst those at risk.

**Simulation study.**

- Focus on estimation of $\alpha$.
- Considered bias as a function of measurement error, censoring rate

Figure 3: Impact of measurement error

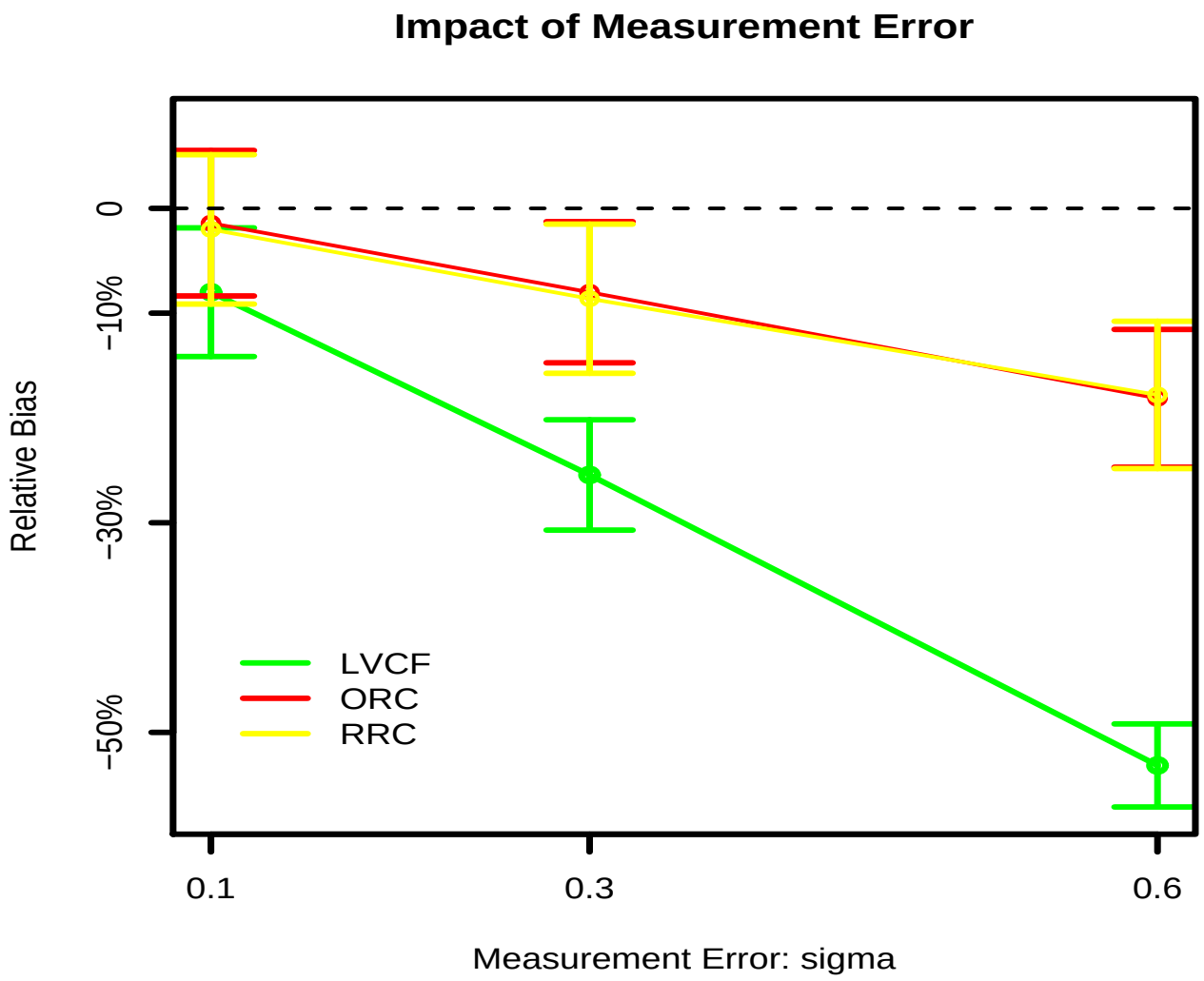
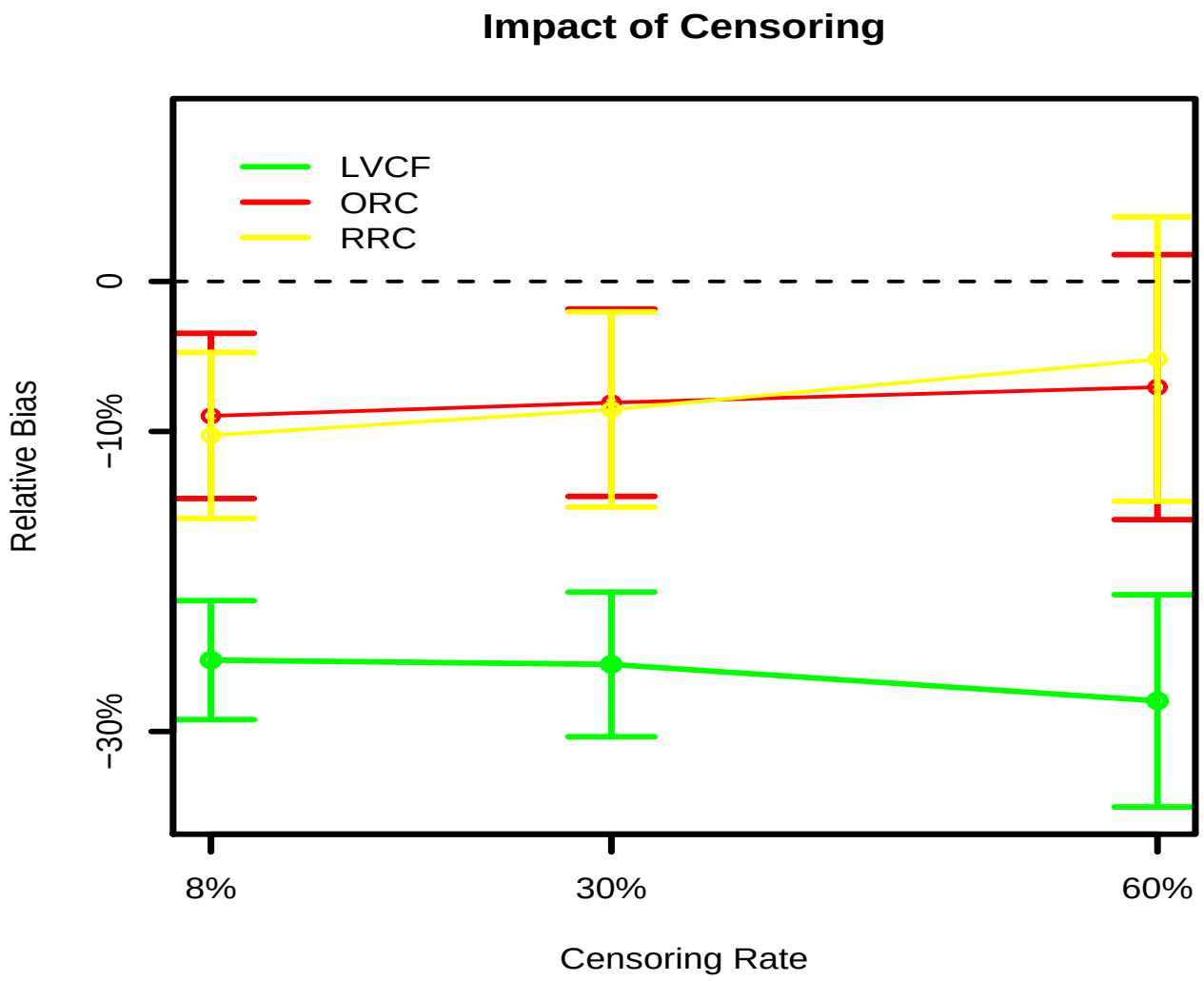


Figure 4: Impact of censoring rate



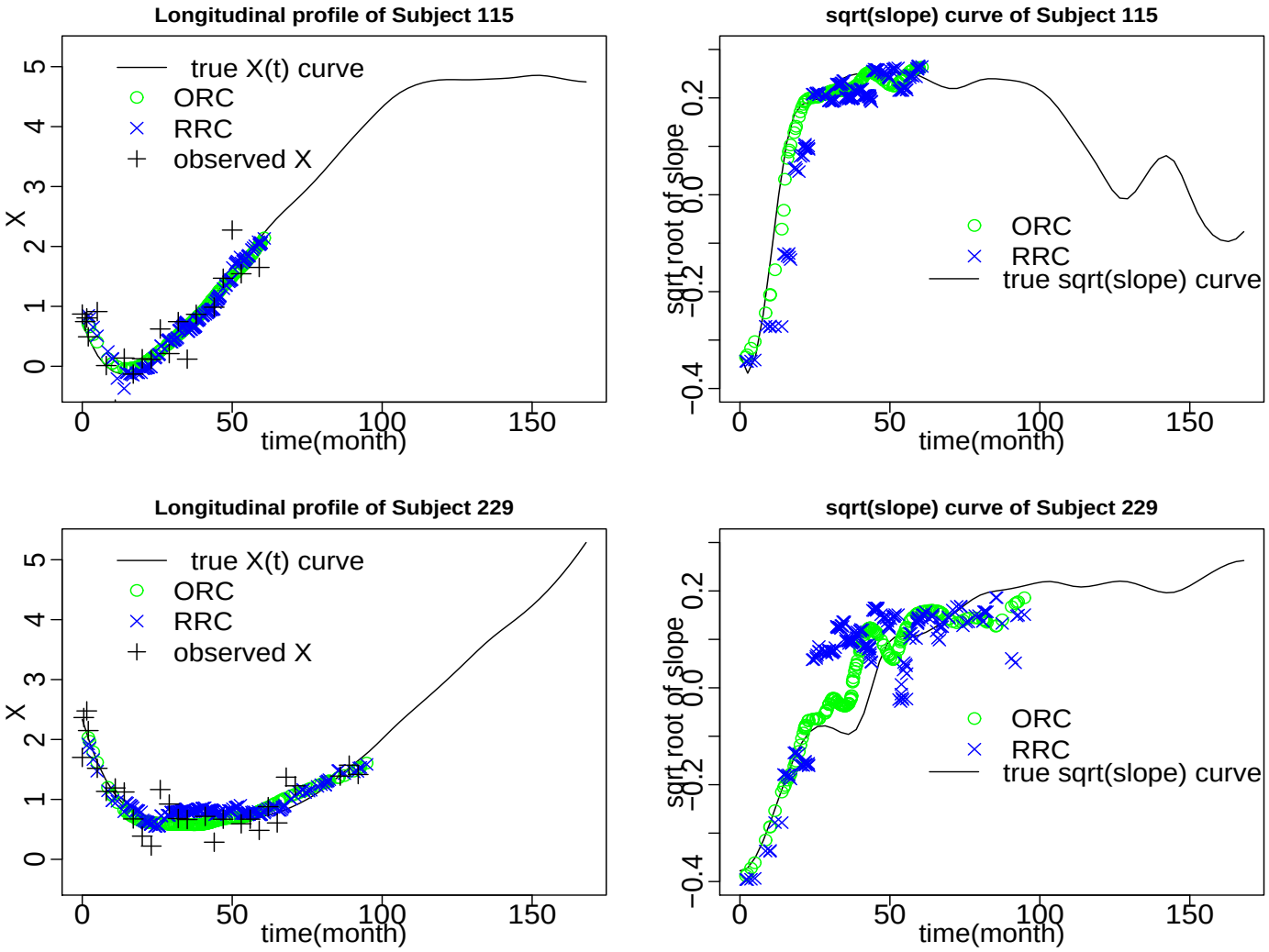
## Conclusion from simulation

- ORC and RRC very similar in this semi-parametric model
- Remaining bias due to two-stage estimation, could be reduced by joint estimation

## Simulated prostate cancer like data

How good is the estimate of the slope from this semi-parametric model?

# Figure 5: Fit to longitudinal data



## Real prostate cancer data

Fit semi-parametric longitudinal model

Figure 6: Fit to longitudinal data

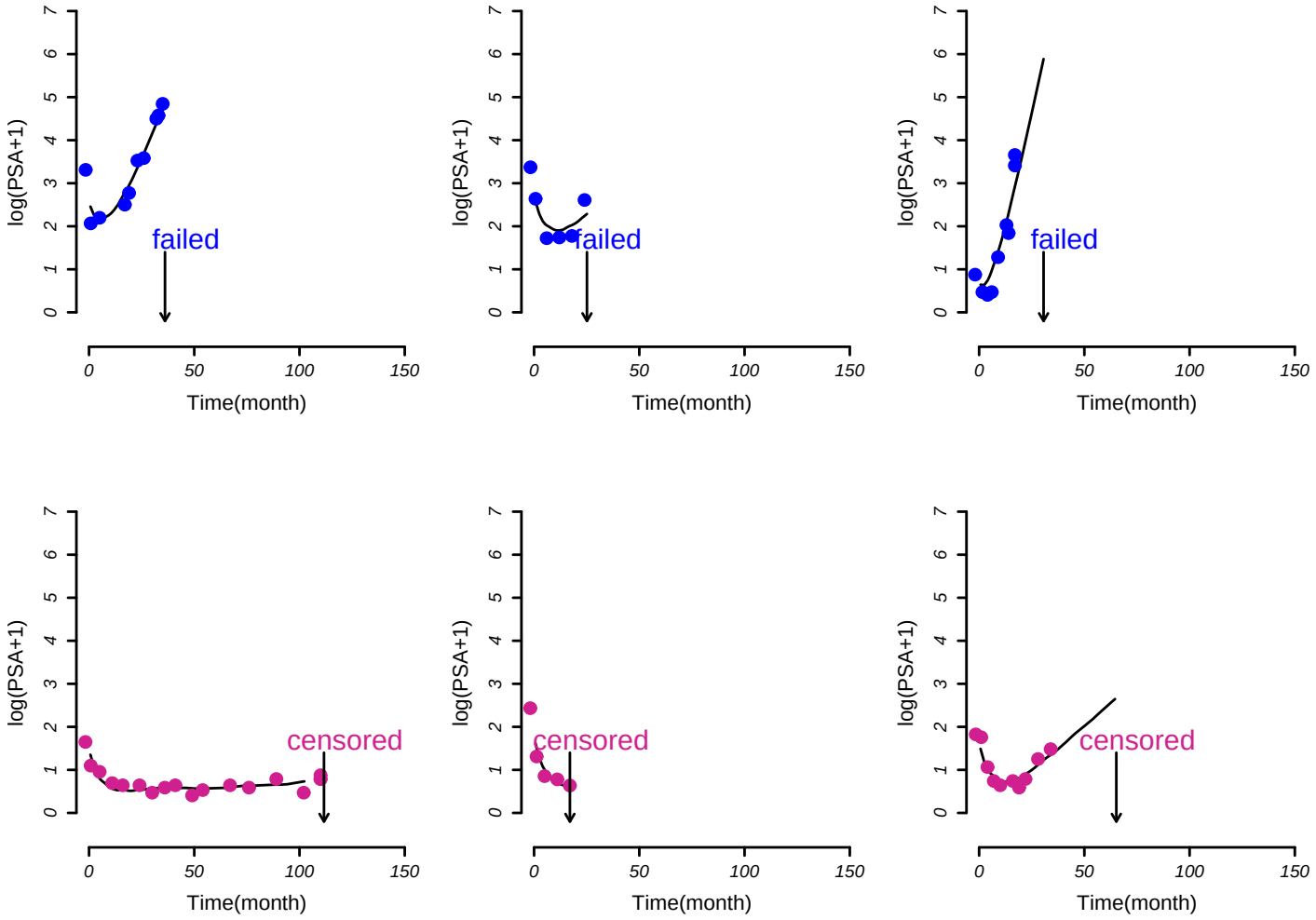
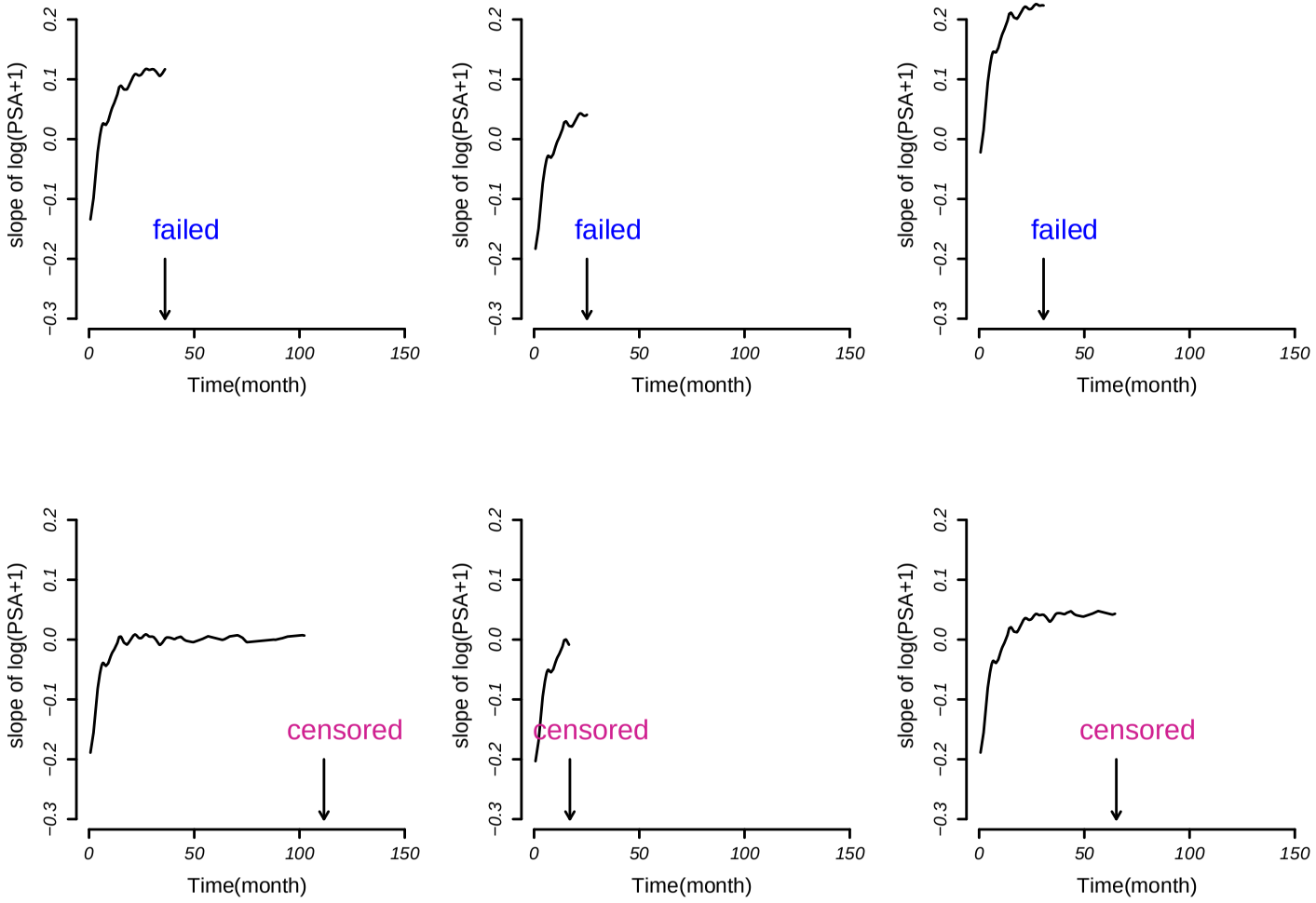


Figure 7: Estimated slope of longitudinal variable



# Open issues

- Assessing model fit
- Robustness issues
- Multivariate longitudinal data
- Non Gaussian longitudinal data
- Efficient algorithms