Penalized Regression Calibration: a statistical method to predict survival from high-dimensional longitudinal covariates

Mirko Signorelli[1]

🏠: mirkosignorelli.github.io
✉: m.signorelli@math.leidenuniv.nl
🐦: @signormirko

Joint work with Roula Tsonaka[2] and Pietro Spitali[2]

[1] Mathematical Institute, Leiden University (NL)
[2] Leiden University Medical Center (NL)

31st International Biometric Conference (IBC2022)
July 11, 2022

Universiteit Leiden

1. Introduction
2. Penalized regression calibration
3. Cluster bootstrap optimism correction
4. R package
5. Applications
6. Conclusion

Appendix

Methodological problem

- High-dimensional & longitudinal covariates more and more common in survival analysis
- high-dimensionality:
    - penalized Cox model, random survival forest...
    - until recently: extensions for longitudinal covs lacking
- Longitudinal covariates → joint models (Rizopoulos, 2012)
    - shared random effects model ⇒ computationally intensive
    - applicability limited to a handful of covariates (Hickey et al., 2016)

⇒ how to predict survival when you have
a high-dimensional set of longitudinal predictors?

Motivating example: predicting time to LoA

The disease: Duchenne muscular dystrophy (DMD)

Rare neuromuscular disorder that leads to

- loss of ambulation (LoA) during adolescence
- premature death (avg: 26 yo)

Motivating example: predicting time to LoA

The disease: Duchenne muscular dystrophy (DMD)

Rare neuromuscular disorder that leads to

- loss of ambulation (LoA) during adolescence
- premature death (avg: 26 yo)

Motivating dataset: MARK-MD study

Observational study on DMD patients (Signorelli et al., 2020)

- $n = 93$ patients ambulant at beginning of study
- up to 5 longitudinal blood samples / patient
- 118 proteins measured using 240 antibodies

Motivating example: predicting time to LoA

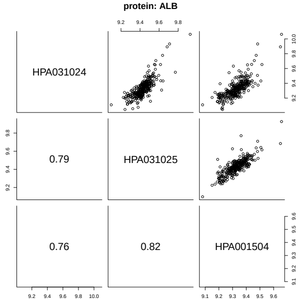
Clinical question:

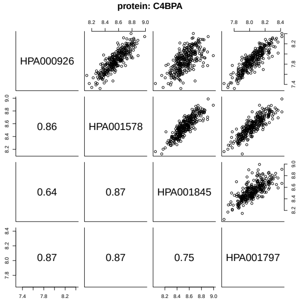
can we use the 240 longitudinal blood biomarkers
to predict time to loss of ambulation in DMD patients?

Challenges in the Mark-MD dataset

Methodological challenges of MarkMD dataset:

1. longitudinal covariates
2. high-dimensionality: 240 antibodies vs 93 patients
3. antibodies targeting same protein are strongly correlated


protein: ALB


protein: C4BPA

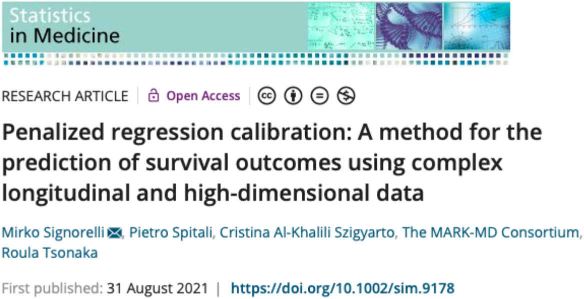
1. Introduction
2. Penalized regression calibration
3. Cluster bootstrap optimism correction
4. R package
5. Applications
6. Conclusion

Appendix

Penalized regression calibration (PRC)

Our solution: penalized regression calibration (Signorelli et al., 2021)

- statistical method to predict survival from high-dimensional longitudinal covariates
- additionally: it can handle groups of strongly correlated predictors



Statistics in Medicine

RESEARCH ARTICLE | Open Access

Penalized regression calibration: A method for the prediction of survival outcomes using complex longitudinal and high-dimensional data

Mirko Signorelli, Pietro Spitali, Cristina Al-Khalili Szigyarto, The MARK-MD Consortium, Roula Tsonaka

First published: 31 August 2021 | https://doi.org/10.1002/sim.9178

SiM article: bit.ly/penregrcal   R package: pencal   🐦: @signormirko

Notation

We consider a setup with:

1. $s = 1, ..., p$ latent biological processes (= proteins)
2. $q = 1, ..., r_s$ items (= antibodies) used to measure protein s
3. $i \in \{1, ..., n\}$ subjects
4. $j = 1, ..., m_i$ repeated measurements for subject i

Notation

We consider a setup with:

1. $s = 1, ..., p$ latent biological processes (= proteins)
2. $q = 1, ..., r_s$ items (= antibodies) used to measure protein s
3. $i \in \{1, ..., n\}$ subjects
4. $j = 1, ..., m_i$ repeated measurements for subject i

Relevant variables:

1. y_{qsij} = level of protein s as measured by antibody q on subject i at visit j

Notation

We consider a setup with:

1. $s = 1, ..., p$ latent biological processes (= proteins)
2. $q = 1, ..., r_s$ items (= antibodies) used to measure protein s
3. $i \in \{1, ..., n\}$ subjects
4. $j = 1, ..., m_i$ repeated measurements for subject i

Relevant variables:

1. y_{qsij} = level of protein s as measured by antibody q on subject i at visit j
2. a_{ij} = age of subject i at visit j

Notation

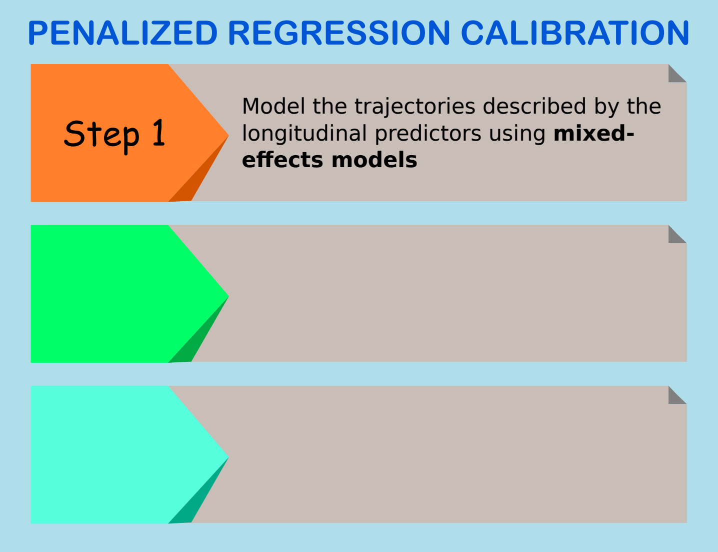
We consider a setup with:

1. $s = 1, ..., p$ latent biological processes (= proteins)
2. $q = 1, ..., r_s$ items (= antibodies) used to measure protein s
3. $i \in \{1, ..., n\}$ subjects
4. $j = 1, ..., m_i$ repeated measurements for subject i

Relevant variables:

1. y_{qsij} = level of protein s as measured by antibody q on subject i at visit j
2. a_{ij} = age of subject i at visit j
3. (t_i, δ_i), where t_i = survival time (from a_{i1}), $\delta_i = 1$ if actual t_i observed, $\delta_i = 0$ if censored

Step 1: model the longitudinal trajectories



PENALIZED REGRESSION CALIBRATION

Step 1

Model the trajectories described by the longitudinal predictors using **mixed-effects models**

Step 1: model for longitudinal biomarkers

- We employ random effects models to describe the longitudinal trajectories of biomarkers
- Two alternative approaches:
  1. Linear Mixed Model (LMM)
  2. Multivariate Latent Process Mixed Model (MLPMM, Proust-Lima et al. (2013))

Univariate approach: LMM

Fit to each antibody $\mathbf{y_{qs}}$ a LMM with random intercept and slope:

$$y_{qsij} = \beta_{qs0} + b_{qs0i} + (\beta_{qs1} + b_{qs1i})a_{ij} + \varepsilon_{qsij},$$

where $b_{qsi} = (b_{qs0i}, b_{qs1i}) \sim N(0, D_{qs})$ and $\varepsilon_{qsi} \sim N(0, \sigma^2_{qs} I_{m_i})$

Univariate approach: LMM

Fit to each antibody $\mathbf{y_{qs}}$ a LMM with random intercept and slope:

$$y_{qsij} = \beta_{qs0} + b_{qs0i} + (\beta_{qs1} + b_{qs1i})a_{ij} + \varepsilon_{qsij},$$

where $b_{qsi} = (b_{qs0i}, b_{qs1i}) \sim N(0, D_{qs})$ and $\varepsilon_{qsi} \sim N(0, \sigma^2_{qs} I_{m_i})$

- ▶ Limitation: LMM approach ignores correlations between antibodies measuring same protein

Multivariate approach: MLPMM

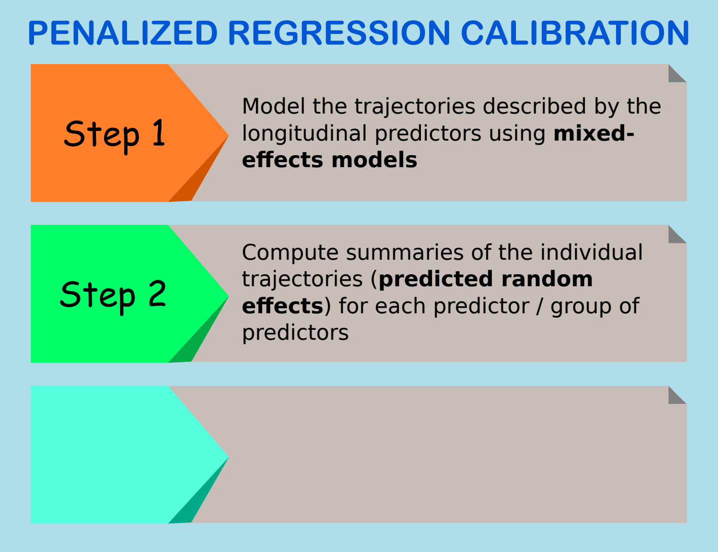
We specify a MLPMM for all antibodies $(\mathbf{y_{1s}}, ..., \mathbf{y_{r_s s}})$ matching protein s:

$$y_{qsij} = \beta_{qs0} + u_{s0i} + b_{qsi} + (\beta_{qs1} + u_{s1i})a_{ij} + \varepsilon_{qsij} \;\; (\forall q = 1, ..., r_s),$$

where $\varepsilon_{qsij} \sim N_1(0, \sigma^2_{\varepsilon qs})$, and

- $\mathbf{u}_{si} = (u_{s0i}, u_{s1i}) \sim N_2(0, \Sigma_{us})$: **shared** (protein-specific) random intercept and slope
- $b_{qsi} \sim N_1(0, \sigma^2_{bqs})$ antibody-specific random intercept, $q = 1, ..., r_s$

Step 2: compute subject-specific summaries



PENALIZED REGRESSION CALIBRATION

Step 1

Model the trajectories described by the longitudinal predictors using **mixed-effects models**

Step 2

Compute summaries of the individual trajectories (**predicted random effects**) for each predictor / group of predictors

Computing the predicted random effects

Derive from the mixed model individual summaries of

- biomarker's heterogeneity → random intercepts
- biomarker's progression rates → random slopes

Computing the predicted random effects

Derive from the mixed model individual summaries of

- biomarker's heterogeneity → random intercepts
- biomarker's progression rates → random slopes

For the LMM:

$$\hat{b}_{qsi} = E(b_{qsi} | Y_{qsi} = y_{qsi}) = D_{qs} Z_i^T V_{qsi}^{-1} (y_{qsi} - X_i \beta_{qs}),$$

where $V_{qsi} = Z_i D_{qs} Z_i^T + \sigma^2_{qs} I_{m_i}$ is the marginal covariance matrix of subject i

Computing the predicted random effects

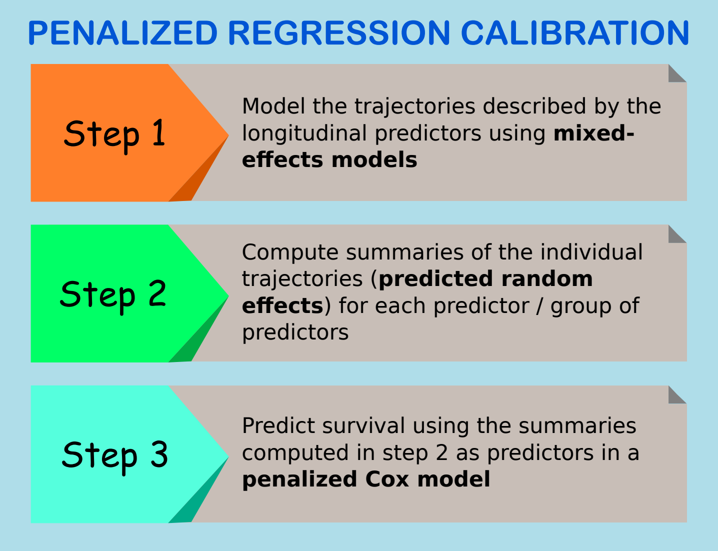
For the MLPMM:

$$\left(\hat{u}_{si}, \hat{b}_{si}\right) = E\left(u_{si}, b_{si} \middle| Y_{si} = y_{si}\right) = \begin{bmatrix} Z_i \Sigma_{u_s} \\ \Sigma_{b_s} I_{r_s} \otimes \mathbb{1}_{m_i,1} \end{bmatrix} \Sigma_{y_{si}}^{-1} \dot{y}_{si},$$

where $y_{si} = (y_{1si1}, ..., y_{1sim_i}, ..., y_{r_s si1}, ..., y_{r_s sim_i})^T$, $\dot{y}_{si}$ is the equivalent of y_{si} with $\dot{y}_{qsij} = y_{qsij} - \beta_{qs0} - \beta_{qs1} a_{ij}$ as entries, Z_i is the random-effects design matrix associated to y_{si}, $\Sigma_{b_s} = \begin{bmatrix} \sigma^2_{b1s} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2_{br_s s} \end{bmatrix}$,

$\Sigma_{\varepsilon_s} = \begin{bmatrix} \sigma^2_{\varepsilon 1s} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2_{\varepsilon r_s s} \end{bmatrix}$, $\Sigma_{u_s} = \begin{bmatrix} \sigma^2_{us0} & \sigma_{us0,us1} \\ \sigma_{us0,us1} & \sigma^2_{us1} \end{bmatrix}$ and
$\Sigma_{y_{si}} = Z_i \Sigma_{us} Z_i^T + I_{r_s} \otimes \Sigma_{\varepsilon s} I_{m_i} + I_{r_s} \otimes \Sigma_{bs} \mathbb{1}_{m_i,m_i}$

Step 3: predict the survival time



PENALIZED REGRESSION CALIBRATION

Step 1

Model the trajectories described by the longitudinal predictors using **mixed-effects models**

Step 2

Compute summaries of the individual trajectories (**predicted random effects**) for each predictor / group of predictors

Step 3

Predict survival using the summaries computed in step 2 as predictors in a **penalized Cox model**

Step 3: model to predict T

- Model for the survival outcome:

$$h(t_i|x_i, \hat{\mathbf{u}}_i, \hat{\mathbf{b}}_i) = h_0(t_i)\exp(\eta_i) \tag{1}$$

- η_i includes baseline covs x_i & predicted random effects. Examples:
  1. PRC LMM: $\eta_i = \xi x_i + \sum_s \sum_q \gamma_{qs}\hat{b}_{0qsi} + \sum_s \sum_q \delta_{qs}\hat{b}_{1qsi}$
  2. PRC MLPMM with $\hat{\mathbf{u}}$: $\eta_i = \xi x_i + \sum_s \gamma_s \hat{u}_{s0i} + \sum_s \delta_s \hat{u}_{s1i}$
  3. PRC MLPMM with $\hat{\mathbf{u}}$ & $\hat{\mathbf{b}}$:
$$\eta_i = \xi x_i + \sum_s \gamma_s \hat{u}_{s0i} + \sum_s \delta_s \hat{u}_{s1i} + \sum_s \sum_q \psi_{qs}\hat{b}_{qsi}$$

Step 3: model to predict T

- Model for the survival outcome:

$$h(t_i | x_i, \hat{\mathbf{u}}_i, \hat{\mathbf{b}}_i) = h_0(t_i) \exp(\eta_i) \tag{1}$$

- Model (1) is high-dimensional ⇒ we estimate it using penalized maximum likelihood

$$\max_{\xi, \gamma, \delta} \; \ell(\xi, \gamma, \delta) - \lambda p(\xi, \gamma, \delta; \alpha)$$

- Penalty functions: ridge (ℓ^2, recommended), elastic net, lasso (ℓ^1)
- Predicted survival: $S_i\hat{(}t) = e^{-\int_0^t \hat{h}_0(s) e^{\hat{\eta}_i} ds}$

1. Introduction

2. Penalized regression calibration

3. Cluster bootstrap optimism correction

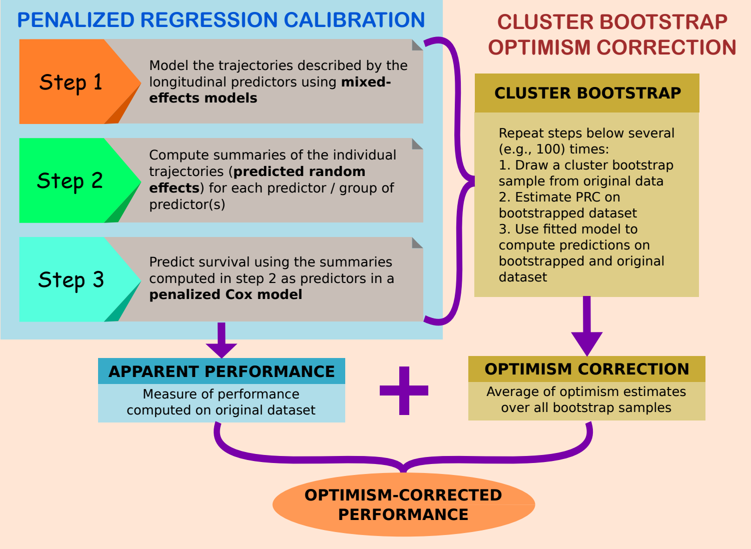
4. R package

5. Applications

6. Conclusion

Appendix

Cluster bootstrap optimism correction procedure



PENALIZED REGRESSION CALIBRATION

Step 1: Model the trajectories described by the longitudinal predictors using **mixed-effects models**

Step 2: Compute summaries of the individual trajectories (**predicted random effects**) for each predictor / group of predictor(s)

Step 3: Predict survival using the summaries computed in step 2 as predictors in a **penalized Cox model**

CLUSTER BOOTSTRAP OPTIMISM CORRECTION

CLUSTER BOOTSTRAP

Repeat steps below several (e.g., 100) times:
1. Draw a cluster bootstrap sample from original data
2. Estimate PRC on bootstrapped dataset
3. Use fitted model to compute predictions on bootstrapped and original dataset

APPARENT PERFORMANCE

Measure of performance computed on original dataset

+

OPTIMISM CORRECTION

Average of optimism estimates over all bootstrap samples

OPTIMISM-CORRECTED PERFORMANCE

1. Introduction

2. Penalized regression calibration

3. Cluster bootstrap optimism correction

4. R package

5. Applications

6. Conclusion

Appendix

The R package pencal

PRC implemented in the R package pencal

- published on CRAN
- both PRC and CBOCP implemented
- optimized for parallel computing
- vignette illustrating how to use the package available on CRAN

1. Introduction

2. Penalized regression calibration

3. Cluster bootstrap optimism correction

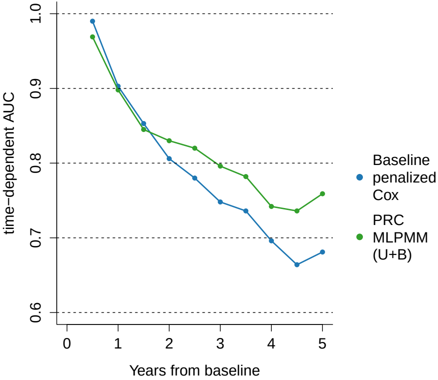
4. R package

5. Applications

6. Conclusion

Appendix

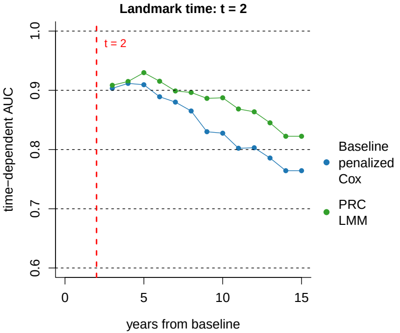
Prediction of time to LoA for DMD patients

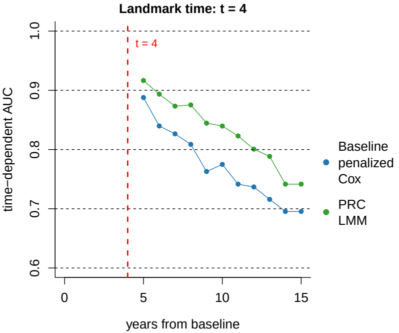


- Exploiting longitudinal information improves predictive performance from $t \geq 2$
- Limitations:
  1. small n (DMD is a rare disease!)
  2. 45 patients with only 1 measurement before LoA ⇒ predicting random slopes challenging

Predicting time to dementia in elderly individuals

- Longitudinal study with follow-up info up to 15 years
- Outcome: time to dementia
- $n = 1634$; many repeated measurements / patient





1. Introduction

2. Penalized regression calibration

3. Cluster bootstrap optimism correction

4. R package

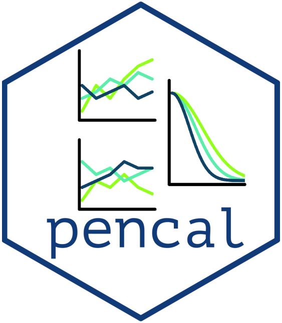
5. Applications

6. Conclusion

Appendix

Take-home message

- PRC makes it possible to predict survival using predictors that are both longitudinal AND high-dimensional
- Idea: if biomarkers' progression rates are associated with T, PRC can improve predictive performance
- Methodology: Signorelli et al. (2021, Statistics in Medicine)
- R package: pencal (available from CRAN)
- Future extensions (interested? Please get in touch! ☺):
  1. GLMMs in step 1
  2. competing risks



🏠: mirkosignorelli.github.io
✉: m.signorelli@math.leidenuniv.nl
🐦: @signormirko

- IBC2022 CoI statement: I have no current or past relationships with commercial entities

References I

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1):117.

Proust-Lima, C., Amieva, H., and Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66(3):470–487.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.

Signorelli, M., Ayoglu, B., Johansson, C., Lochmüller, H., Straub, V., Muntoni, F., Niks, E., Tsonaka, R., Person, A., Aartsma-Rus, A., Nilsson, P., Al-Khalili Szigyarto, C., and Spitali, P. (2020). Longitudinal serum biomarker screening identifies MDH2 as candidate prognostic biomarker for Duchenne muscular dystrophy. *Journal of Cachexia, Sarcopenia and Muscle*, 11(2):505–517.

Signorelli, M., Spitali, P., Al-Khalili Sgyziarto, C., The Mark-MD Consortium, and Tsonaka, R. (2021). Penalized regression calibration: a method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*.

1. Introduction

2. Penalized regression calibration

3. Cluster bootstrap optimism correction

4. R package

5. Applications

6. Conclusion

Appendix

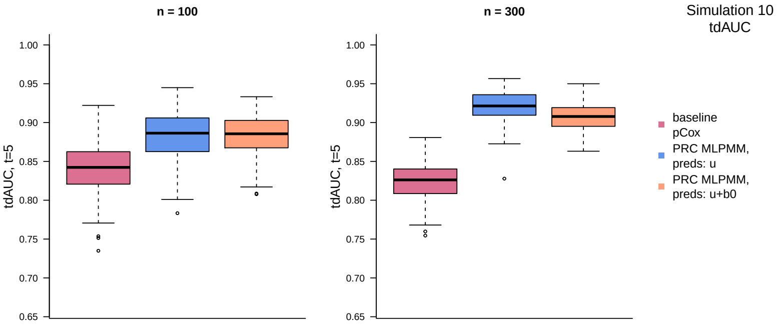
Simulation settings

- $n = 100$ & $n = 300$
- y_{qs}: $p = 50$ proteins, each with $r_s = 3$ antibodies
- $T \rightarrow$ Weibull model
- simulation 10: T depends on shared random effects u_{s0}, u_{1s} only, not on item-specific b_{qs}[1]
- models compared:
  1. penalized Cox with baseline measurements
  2. PRC-MLPMM(U)
  3. PRC-MLPMM(U+B)

[1] See Signorelli et al. (2021) for more simulations

Model performance (time-dependent AUC at $t = 5$)

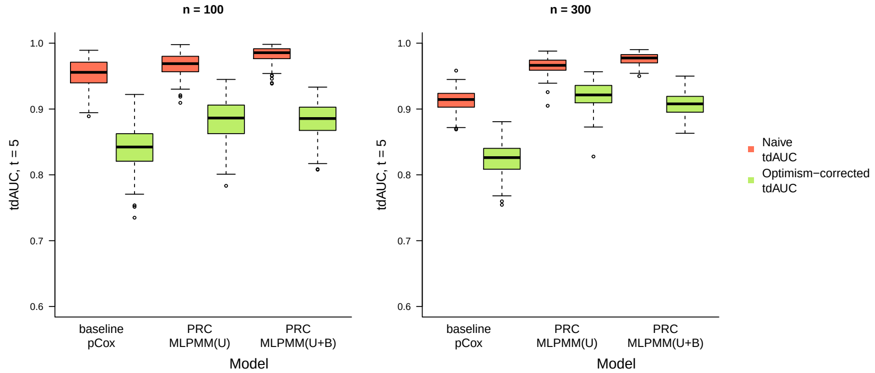
Simulation 10: $T = f(u_0, u_1)$



- improvement of PRC stronger when n larger: mixed models yield better summaries + lower variability of performance measures

Effect of the CBOCP

Naive vs optimism-corrected tdAUC in simulation 10:



- CBOCP needed to avoid reporting optimistic performance
- Issue particularly important with small *n*