Pseudo-values approach for Quantile analysis in Individual Patient Data Meta-analysis

Alessandra Meddis*, O. Bouaziz, X.Paoletti, A.Latouche

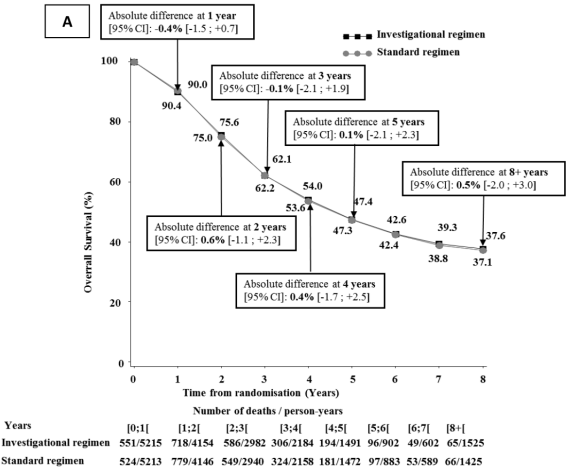
*Department of Public Health, University of Copenhagen, Denmark

ISCB44
August 28 2023, Milan

KØBENHAVNS UNIVERSITET

Motivation

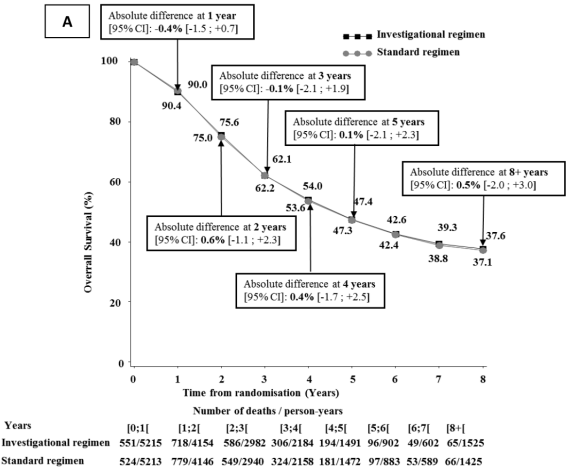
Individual Patient Data (IPD) meta-analysis for first-line therapy in advanced ovarian cancer



- 11029 women from 17 clinical trials
- Individual characteristics: age,tumor stage, tumor grade, surgical procedures,...
- Previous analysis:
  - HRs by two-stage fixed-effect approach
  - Peto curve: pooled trial-specific Log-rank statistics at each year

Motivation

Individual Patient Data (IPD) meta-analysis for first-line therapy in advanced ovarian cancer



- 11029 women from 17 clinical trials
- Individual characteristics: age,tumor stage, tumor grade, surgical procedures,...
- Current analysis:
  - treatment benefit, for fixed quantiles (difference in survival times).
  - one-step approach

PSEUDO-VALUE APPROACH FOR CONDITIONAL QUANTILE RESIDUAL LIFETIME ANALYSIS FOR CLUSTERED SURVIVAL AND COMPETING RISKS DATA WITH APPLICATIONS TO BONE MARROW TRANSPLANT DATA

Kwang Woo Ahn and **Brent R. Logan**
Medical College of Wisconsin

Abstract

Quantile residual lifetime analysis is conducted to compare remaining lifetimes among groups for survival data. Evaluating residual lifetimes among groups after adjustment for covariates is often of interest. The current literature is limited to comparing two groups for independent data. We propose a pseudo-value approach to compare quantile residual lifetimes given covariates between multiple groups for independent and clustered survival data. The proposed method considers clustered event times and clustered censoring times in addition to independent event times and censoring times. We show that the method can also be used to compare multiple groups on the cause specific residual life distribution in the competing risk setting, for which there are no current methods which account for clustering. The empirical Type I errors and statistical power of the proposed study are examined in a simulation study, which shows that the proposed method controls Type I errors very well and has higher power than an existing method. The proposed method is illustrated by a bone marrow transplant data set.

Ahn and Logan

"...We propose a pseudo-value approach to compare quantile residual lifetimes given covariates between multiple groups for independent and clustered survival data..."

1. Extend pseudo-values to clustered data (individual i in cluster k)

$$P_{ik}(t) = n(\hat{S}(t)) - (n-1)\hat{S}^{-ik}(t), \quad i = 1, ..., n_k \quad k = 1, ..., m$$

   with $\hat{S}(t)$ the Kaplan-Maier estimate on the pooled data.
2. For a fixed τ^*, calculate $\hat{q}_{t^*}$ such that: $\hat{S}(\hat{q}_{t^*}|z_0) = 1 - \tau^*$ by Cox regression.
3. Fit a GEE on the pseudo-values at the specific quantile $P_{ik}(\hat{q}^*_t)$:

$$\lim_{m \to \infty} E[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$$

4. Propose a test to compare the conditional residual quantiles between multiple groups based on the predicted values of the GEE and the estimated covariance matrix

Ahn and Logan

"...We propose a pseudo-value approach to compare quantile residual lifetimes given covariates between multiple groups for independent and clustered survival data..."

1. $\lim_{m\to\infty} E[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$ (NOT FOR Meta-Analysis)
2. Propose a **test** to compare the conditional residual quantiles between multiple groups based on the predicted values of the GEE and the estimated covariance matrix

Our proposal

1. Extend pseudo-values to meta-analysis
2. Propose an estimate for the difference in survival time at a fixed quantile
   - Quantile : more flexible and robust quantitative tools for mean-based regression models. Moreover, the proportional hazard assumption is not needed and it allows us to detect potential late treatment effects

Notations

- m clusters with sample size n_k , $n = \sum_{k=1}^{m} n_k$
- T_{ik}, C_{ik} event time and censoring time
- $Z_{ik} = \{X_{ik}, Z_{ik}^1, ..., Z_{ik}^p\}$ covariate for individual i in cluster k, with X_{ik} treatment covariate

We assume:

- $n_k \gg m$ $(m < 20)$
- Independent censoring
- Event times might be correlated within the same cluster

Our proposal

1. Extend pseudo-values to meta-analysis:

$$P_{ik}(t) = n\tilde{S}(t) - (n-1)\tilde{S}^{-ik}(t) \quad i = 1, ..., n_k, \quad k = 1, ..., m$$

with

$$\tilde{S}(t) = \frac{\sum_{k=1}^{K} w_k \hat{S}_k(t)}{\sum_{k=1}^{K} w_k}, \quad w_k = 1/(\sigma_k^2 + \tau^2)$$

where σ_k^2 is the variance for the trial-specific survival function estimate by Kaplan-Meier and τ^2 represents the between-trial variability[1] .

[1] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177–88.

Pseudo-values

$$P_{ik}(t) = n\tilde{S}(t) - (n-1)\tilde{S}^{-ik}(t) \quad i = 1, ..., n_k, \quad k = 1, ..., m$$

Proposition:

$\lim_{n\to\infty} \mathrm{E}[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$ if $\lim_{n_k\to\infty} \mathrm{E}[P^k_{ik}(t)|Z_{ik}] = S_k(t|Z_{ik}) \quad \forall k$

where $P^k_{ik}(t) := n_k\hat{S}_k(t) - (n_k - 1)\hat{S}^{-ik}_k(t)$

Idea of Proof

We use the decomposition $\sum_k w_k\hat{S}^{-ij}_k(t) = \sum_{k\neq j} w_k\hat{S}_k(t) + w_j\hat{S}^{-ij}_j(t)$. With some algebra:

$$P_{ij}(t) = ... = \sum_k w_k\hat{S}_k(t) + w_j[-\hat{S}_j(t) + P^j_{ij}(t)] + (n - n_j)w_j[\hat{S}_j(t) - \hat{S}^{-ij}_j(t)]$$

Pseudo-values

$$P_{ik}(t) = n\tilde{S}(t) - (n-1)\tilde{S}^{-ik}(t) \quad i = 1, ..., n_k, \quad k = 1, ..., m$$

Proposition:

$\lim_{n\to\infty} \mathrm{E}[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$ if $\lim_{n_k\to\infty} \mathrm{E}[P^k_{ik}(t)|Z_{ik}] = S_k(t|Z_{ik}) \quad \forall k$

where $P^k_{ik}(t) := n_k\hat{S}_k(t) - (n_k - 1)\hat{S}_k^{-ik}(t)$

Idea of Proof

when $n_j \to \infty$

$$\mathrm{E}[P_{ij}(t)|Z_{ij}] = \mathrm{E}[\tilde{S}(t)|Z_{ij}] + w_j \underbrace{\mathrm{E}[-\hat{S}_j(t) + P^j_{ij}(t)|Z_{ij}]}_{0} + (n - n_j)w_j\mathrm{E}[\hat{S}_j(t) - \hat{S}_j^{-ij}(t)|Z_{ij}]$$

Pseudo-values

$$P_{ik}(t) = n\tilde{S}(t) - (n-1)\tilde{S}^{-ik}(t) \quad i = 1, ..., n_k, \quad k = 1, ..., m$$

Proposition:

$\lim_{n\to\infty} \mathrm{E}[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$ if $\lim_{n_k\to\infty} \mathrm{E}[P^k_{ik}(t)|Z_{ik}] = S_k(t|Z_{ik}) \quad \forall k$

where $P^k_{ik}(t) := n_k\hat{S}_k(t) - (n_k - 1)\hat{S}_k^{-ik}(t)$

Idea of Proof

when $n_j \to \infty$

$$\mathrm{E}[P_{ij}(t)|Z_{ij}] = \mathrm{E}[\tilde{S}(t)|Z_{ij}] + w_j\mathrm{E}[-\hat{S}_j(t) + P^j_{ij}(t)|Z_{ij}] + (n - n_j)w_j \underbrace{\mathrm{E}[\hat{S}_j(t) - \hat{S}_j^{-ij}(t)|Z_{ij}]}_{0}$$

Pseudo-values

$$P_{ik}(t) = n\tilde{S}(t) - (n-1)\tilde{S}^{-ik}(t) \quad i = 1, ..., n_k, \quad k = 1, ..., m$$

Proposition:

$\lim_{n\to\infty} \mathsf{E}[P_{ik}(t)|Z_{ik}] = S(t|Z_{ik})$ if $\lim_{n_k\to\infty} \mathsf{E}[P^k_{ik}(t)|Z_{ik}] = S_k(t|Z_{ik}) \quad \forall k$

where $P^k_{ik}(t) := n_k\hat{S}_k(t) - (n_k - 1)\hat{S}^{-ik}_k(t)$

Idea of Proof

when $n_j \to \infty$

$$\mathsf{E}[P_{ij}(t)|Z_{ij}] = \underbrace{\mathsf{E}[\tilde{S}(t)|Z_{ij}]}_{S(t|Z_{ij})^2} + w_j\mathsf{E}[-\hat{S}_j(t) + P^j_{ij}(t)|Z_{ij}] + (n - n_j)w_j\mathsf{E}[\hat{S}_j(t) - \hat{S}^{-ij}_j(t)|Z_{ij}]$$

[2] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177–88.

Method

- Fix $Z = z_0$ and τ^*, calculate $\hat{q}_{t^*}$ such that: $\hat{S}(\hat{q}_{t^*}|z_0) = 1 - \tau^*$
- Calculate the pseudo-values $P_{ik}(\hat{q}_{t^*})$
- Fit a GEE on the pseudo-values to estimate $S(\hat{q}_{t^*}|z_0)$ with jacknife estimator for the variance:

$$\sqrt{n}(\hat{S}(\hat{q}_{t^*}|z_0) - S(\hat{q}_{t^*}|z_0)) \to \mathcal{N}(0, \Sigma)$$

- We can construct the confidence interval

$$\mathcal{P}[-z_{\frac{\alpha}{2}}\frac{\hat{\Sigma}}{\sqrt{n}} + \hat{S}(\hat{q}_{t^*}|z_0) \leq S(\hat{q}_{t^*}|z_0)) \leq +z_{\frac{\alpha}{2}}\frac{\hat{\Sigma}}{\sqrt{n}} + \hat{S}(\hat{q}_{t^*}|z_0)]$$

$$\mathrm{CI}(\hat{q}_{t^*}) : [S^{-1}(\pm\ z_{\frac{\alpha}{2}}\frac{\hat{\Sigma}}{\sqrt{n}} + \hat{S}(\hat{q}_{t^*}|z_0)\ )\ ]$$

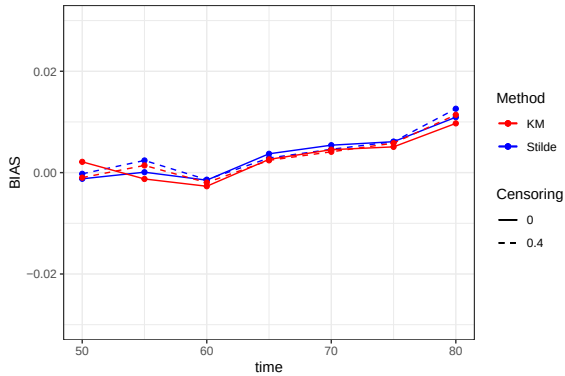
Simulation: Generation of data

$$\lambda_0(t)\exp(b_k + \log(2) * X + 0.15 * Z_1 + 0.3 * Z_2 - 0.1 * Z_3)$$

- $n_k \sim Pois(\lambda = 200)$
- $m = 5, 10 \rightarrow n \sim 1000, 2000$
- Weibull $\lambda_0(t)$
- $b_k \in [0.5, 1.5]$
- $X \sim Bin(0.5)$
- $Z_1 \sim \mathcal{N}(0, 0.4), Z_2 \sim \mathcal{N}(0.5, 0.3), Z_3 \sim Bin(0.5)$
- $C \sim Exp(0.007) \rightarrow 40\%$
- 200 repetitions

Simulation I: Estimate the conditional Survival function by the proposed approach with the DerSimonian estimate for the survival ($\tilde{S}(t)$)

Simulation results Part I



Simulation Part II

$$\lambda_0(t)\exp(b_k + \log(2) * X + 0.15 * Z_1 + 0.3 * Z_2 - 0.1 * Z_3)$$

Estimate the survival time at that specific quantile ($\tau^* = 0.6$) and the confidence interval with the proposed method, where $\hat{q}_t^*$ (step 1) is estimated by

- Random Survival Forest (on the pooled data with trial-specific covariate)
- Marginalized Cox (trial-specific Cox, marginalize respect to the trial distribution)

Two sources of variability:

- Estimation of q_{t^*}
- Estimation of $S(q_{t^*}|_0)$

Simulation results Part II

- Estimation of $S(q_{t^*}|z_0)$:
  - BIAS: $S(q_{t^*}) - \tilde{S}(q_{t^*})$, BIAS1: $S(q_{t^*}) - \hat{S}(\hat{q}_{t^*})$, BIAS2: $\tilde{S}(\hat{q}_{t^*}) - \hat{S}(\hat{q}_{t^*})$

			Random Forest		Marginalized Cox	
m	censoring	Bias	Bias1	Bias2	Bias1	Bias2
5	0%	0.001	0.060	0.056	0.001	0.000
	40%	0.002	0.057	0.053	0.001	-0.005
10	0%	-0.004	0.024	0.020	$< 10^{-3}$	-0.003
	40%	-0.002	0.017	0.013	$< 10^{-3}$	-0.002

Simulation results Part II

- Estimation of q_{t^*}

m	censoring	Random Forest		Marginalized Cox	
		Bias	cov prob	Bias	cov prob
5	0%	2.296	0.790	-0.041	0.890
	40%	2.154	0.755	-0.102	0.855
10	0%	1.081	0.935	0.040	0.925
	40%	0.692	0.940	-0.002	0.910

Conclusions

- Discussion:
  - Method highly depends on the initial estimation of the quantile
  - Coverage probability (improve variance estimation)
- On the Agenda:
  - Extend the simulation to more complex settings
  - Extend the method to the difference of the survival times for 2 or more groups
  - Investigate the case of heterogeneous follow-up
    - How much does impact it?
    - Should we do some extrapolation?
  - Application to the motivational study

References

1. Ahn, Kwang Woo, and Brent R. Logan. "Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data." The annals of applied statistics 10.2 (2016): 618.
2. Paoletti, Xavier, et al. "Assessment of progression-free survival as a surrogate end point of overall survival in first-line treatment of ovarian cancer: a systematic review and meta-analysis." JAMA network open 3.1 (2020): e1918939-e1918939.
3. Ishwaran, Hemant, et al. "Random survival forests." (2008): 841-860.
4. Linear Regression Analysis for Highly Stratified Failure Time Data.

DerSimonian- Laird estimate

$$\tau^2 = \max\left(0, \frac{Q - (m-1)}{S_1 - \frac{S_2}{S_1}}\right)$$

with $S_r = \sum_{k=1}^{m} (1/\sigma_k^2)^r,\ r = 1, 2$ and

$$Q = \sum_{k=1}^{m} \frac{1}{\sigma_k^2} (\hat{S}_k(t) - \hat{S}^{\tau}(t))$$

$$\hat{S}^{\tau}(t) = \frac{\sum_{k=1}^{m} \frac{1}{\sigma_k^2} \hat{S}_k(t)}{\sum_{k=1}^{m} \frac{1}{\sigma_k^2}}$$

Proposed method: illustration