Nonparametric two-sample tests of longitudinal data in the presence of a terminal event

Jinheum Kim[1], Yang-Jin Kim,[2] & Chung Mo Nam[3]

[1]Department of Applied Statistics, University of Suwon,
[2]Department of Statistics, Ewha Womans University,
[3]Department of Preventive Medicine, Yonsei University College of Medicine

August 19, 2009

Contents

- Analysis of longitudinal data
- Marked point process
- Proposed tests
- Simulations
- An example: CSL 1 data set
- Concluding remarks

Longitudinal data

- In longitudinal studies, subjects are measured repeatedly over time
- For subject $i = 1, \ldots, n$, observed data are

$$(T_{ij}, Z_{ij}, \mathbf{x}_{ij}), \quad j = 1, \ldots, k_i,$$

  where T_{ij} is a measurement time (observation time), and Z_{ij} and $\mathbf{x}_{ij}$ are response and a vector of covariates measured at T_{ij}, respectively
- In ordinary longitudinal data,
  - measurement times are fixed
  - k_i are same for all subjects, i.e., $k_i = k, \forall i$
- However, longitudinal studies related with human being are not easy to keep this scheme

Statistical modeling of longitudinal data

- Two objectives for statistical models of longitudinal data
  - To adopt the conventional regression tools, which relate the response variables to the explanatory variables
  - To account for the within-subject correlation
- In most longitudinal studies,
  - the regression objective is of primary interest
  - the within-subject correlation is essential, but is often of secondary interest
- Three modeling approaches with longitudinal data (Diggle, Heagerty, Liang & Zegger, 2002)

Three approaches

- Marginal model approach
  - $\mathrm{E}(Z_{ij}) = \mu_{ij}$ is related to $\mathbf{x}_{ij}$ by

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\beta,$$

    where g is a known link function
  - $\mathrm{Var}(Z_{ij}) = \nu(\mu_{ij})\phi$, where ν is a known function and ϕ is the over-dispersion parameter
  - $\mathrm{Cov}(Z_{ij}, Z_{ik}) = c(\mu_{ij}, \mu_{ik}; \alpha)$, where c is a known function and α is the additional parameter.
- Random effects model approach
  - $g(\mathrm{E}(Z_{ij}|b_i)) = \mathbf{x}'_{ij}\beta^* + \mathbf{w}'_{ij}b_i$, where $\mathbf{w}_{ij}$ is a vector of covariates whose coefficients vary across subjects; $b_i \sim F(\mathbf{\Theta})$ with mean zero and covariance matrix $\mathbf{\Theta}$, where F is a known distribution function
  - Assume the conditional independence of $Z_{i1}, \ldots, Z_{ik_i}$ given b_i

Three approaches

- Transition model approach
  - $\mathrm{E}(Z_{ij}|Z_{i,j-1},\ldots,Z_{i1}) = \mu_{ij}^c$ depends on $\mathbf{x}_{ij}$ and past responses,

$$g(\mu_{ij}^c) = \mathbf{x}_{ij}'\beta^{**} + \sum_k \alpha_k^{**} f_k(Z_{i,j-1},\ldots,Z_{i1}),$$

    where $f_k, k = 1,\ldots$ are known functions
  - $\mathrm{Var}(Z_{ij}|Z_{i,j-1},\ldots,Z_{i1}) = \nu(\mu_{ij}^c)\phi$, where ν is a known function and ϕ is the over-dispersion parameter

Goal

- Dynamic model approach proposed by Scheike & Zhang (1998), Martinussen & Scheike (1999, 2000) among others
  - Describe longitudinal data consisting of the triplet (responses, observation times, covariates) using a marked point process
  - Model the conditional mean of the current response given past outcomes, which amount to previously obtained measurements and the times for these measurements ($\equiv$ transition model approach)
- Propose nonparametric tests whether two groups of longitudinal response data have identical conditional mean functions in the presence of group-specific observation and/or termination times

Marked point process

- Denote $(E, \mathcal{E})$ by a measurable mark space
- Let $(Z_k, k \geq 1)$ be a sequence of rvs in E and the sequence $(T_k, k \geq 1)$ constitutes a counting process $N(t) = \sum_k I(T_k \leq t)$
- The double sequence (T_k, Z_k) is called a marked point process (MPP) with an associated counting process

$$N(A, t) = \sum_k I(Z_k \in A) I(T_k \leq t), A \in \mathcal{E}$$

  Specially, $N(t) = N(E, t)$
- The MPP can be identified with counting measure $p(ds \times dz)$ defined by $p((0, t] \times A) = N(A, t), A \in \mathcal{E}$.
- Let $\mathcal{F}_{T_k-} = \sigma(T_j, Z_j, 1 \leq j \leq k-1; T_k)$

Marked point process

- The marked point counting process $N(A,t)$ has the intensity function

$$\lambda_t(dz) = \lambda(t)\Phi_t(dz),$$

where $\lambda(t)$ is a nonnegative $\mathcal{F}_t-$predictable process and Φ_t is defined as $\Phi_t(A) = \Pr(Z(t) \in A|\mathcal{F}_{t^-})$

- For a $\mathcal{F}_t-$predictable process H, the MPP integral

$$\int_0^t \int_E H(s,z)p(ds \times dz)$$

may be decomposed as

$$\int_0^t \lambda(s)\{\int_E H(s,z)\Phi_s(dz)\}ds + \int_0^t \int_E H(s,z)q(ds \times dz) \qquad (1)$$

where $q(dt \times dz) = p(dt \times dz) - \lambda(t)\Phi_t(dz)dt$ is a marked point martingale.

Data structure & notations

- Let $(R, \mathcal{B})$ denote a mark space, where $\mathcal{B}$ is the Borel σ-field on $R = (-\infty, \infty)$
- Let $D_{k,i}$ denote the death time and $C_{k,i}$ the censoring time, where $k = 1, 2; i = 1, \ldots, n_k$
- Assume that $C_{k,i}$ is independent of both $D_{k,i}$ and $N_{k,i}(\cdot, \cdot)$
- Due to censoring, $D_{k,i}$ and $N_{k,i}(\cdot, \cdot)$ may not be fully observed
- For $A \in \mathcal{B}$, we observe

$$\{(N_{k,i}(A, \cdot \wedge C_{k,i}), Z^{k,i}(\cdot \wedge C_{k,i}), X_{k,i}, \delta_{k,i}) | k = 1, 2; i = 1, \ldots, n_k\},$$

where $X_{k,i} = D_{k,i} \wedge C_{k,i}$, and $\delta_{k,i} = I(D_{k,i} \leq C_{k,i})$.

Data structure & notations

- Let $Y_{k,i}(t) = I(X_{k,i} \geq t)$ and $Y_{k\cdot}(t) = \sum_{i=1}^{n_k} Y_{k,i}(t)$
- For the longitudinal responses, *i.e.* marks, it is modeled as

$$Z^{k,i}(t) = m_k(t) + \epsilon_{k,i}, \tag{2}$$

where $m_k(t)$ is a smooth mean function and $\epsilon_{k,i}$ has mean zero and variance σ_k^2

- For the observation times, the intensity process $\lambda_{k,i}(t)$ can be written as

$$\lambda_{k,i}(t) = \alpha_k(t) Y_{k,i}(t), \tag{3}$$

where $\alpha_k(t)$ is a deterministic function given the accrued information up to time $t-$

Cumulative mean function

- Let $\mu_k(t) = \int_0^t m_k(s)ds$ be the cumulative mean function for group k
- Under the assumed models (2) and (3), for any fixed t, using the decomposition (1), we have the decomposition

$$\int_R z p_{k,i}(dt \times dz) = \alpha_k(t) Y_{k,i}(t) d\mu_k(t) + \int_R z q_{k,i}(dt \times dz)$$

- Estimate $\mu_k(t)$ by

$$\hat{\mu}_k(t) = \sum_{i=1}^{n_k} \int_0^t \int_R J_k(s) Y_{k,i}(s) \frac{z}{\hat{\alpha}_k(s) Y_{k\cdot}(s)} p_{k,i}(ds \times dz),$$

where $J_k(t) = I\{Y_{k\cdot}(t) > 0, \hat{\alpha}_k(t) > 0\}$ and $\hat{\alpha}_k(t)$ is the kernel-smoothed estimate of the Nelson-Aalen estimator of $A_k(t) = \int_0^t \alpha_k(s)ds$

Weighted cumulative mean function

- Define a weighted cumulative mean function as

$$\psi_k(t) = \int_0^t S_k(s) d\mu_k(s),$$

where $S_k(t) = \Pr(D_{k,i} \geq t)$

- Estimate $\psi_k(t)$ by

$$\hat{\psi}_k(t) = \int_0^t \hat{S}_k(s) d\hat{\mu}_k(s),$$

where $\hat{S}_k(t)$ is the Kaplan-Meier estimator of $S_k(t)$ based on $\{(X_{k,i}, \delta_{k,i}) | i = 1, \ldots, n_k\}$

Asymptotic distribution of $\hat{\psi}_k(t)$

- Under the regularity conditions,

$$n_k^{1/2}\{\hat{\psi}_k(t) - \psi_k(t)\} \xrightarrow{d} U_k(t), \quad t \in [0, \tau_k],$$

where $U_k(t)$ is a mean zero Gaussian process whose covariance function at (s, t) consistently estimated by

$$\hat{\xi}_k(s,t) = n_k^{-1} \sum_{i=1}^{n_k} \hat{\Psi}_{k,i}(s)\hat{\Psi}_{k,i}(t),$$

where

$$\hat{\Psi}_{k,i}(t) = \int_0^t \int_R J_k(s) \frac{\hat{S}_k(s)}{\hat{\alpha}_k(s) Y_{k\cdot}(s)/n_k} \{Y_{k,i}(s)z - \hat{m}_k(s)\} p_{k,i}(ds \times dz)$$
$$- \hat{\mu}_k(t) \int_0^t \frac{d\hat{M}^D_{k,i}(s)}{Y_{k\cdot}(s)/n_k} + \int_0^t \hat{\mu}_k(s) \frac{d\hat{M}^D_{k,i}(s)}{Y_{k\cdot}(s)/n_k}$$

Nonparametric two-sample tests

- To compare mean functions of two groups when two groups have both different distributions for the observation times and/or different intensities for the termination time
- Hypothesis of interest: $\mathrm{H}_0 : \psi_1(t) = \psi_2(t), \forall t \in (0, \tau]$, where $\tau = \tau_1 \wedge \tau_2$
- Two test statistics
  - $Q_{\mathrm{C}} = \hat{\psi}_1(\tau) - \hat{\psi}_2(\tau)$
  - $Q_{\mathrm{LR}} = \int_0^\tau \hat{K}_{\mathrm{LR}}(s) d\{\psi_1(s) - \psi_2(s)\}$, where $\hat{K}_{\mathrm{LR}}(t) = \frac{Y_{1\cdot}(t) Y_{2\cdot}(t)}{Y_{1\cdot}(t) + Y_{2\cdot}(t)} \frac{n}{n_1 n_2}$

Asymptotic null distribution

- $(n_1 n_2/n)^{1/2} Q_{\mathrm{C}}$ and $(n_1 n_2/n)^{1/2} Q_{\mathrm{LR}}$ converge in distribution to mean zero normal random variables with variances consistently estimated by

$$\hat{\Sigma}_{\mathrm{C}} = \frac{n_2}{n n_1} \sum_{i=1}^{n_1} \left\{ \int_0^\tau d\hat{\Psi}_{1,i}(s) \right\}^2 + \frac{n_1}{n n_2} \sum_{i=1}^{n_2} \left\{ \int_0^\tau d\hat{\Psi}_{2,i}(s) \right\}^2$$

and

$$\hat{\Sigma}_{\mathrm{LR}} = \frac{n_2}{n n_1} \sum_{i=1}^{n_1} \left\{ \int_0^\tau \hat{K}_{\mathrm{LR}}(s) d\hat{\Psi}_{1,i}(s) \right\}^2 + \frac{n_1}{n n_2} \sum_{i=1}^{n_2} \left\{ \int_0^\tau \hat{K}_{\mathrm{LR}}(s) d\hat{\Psi}_{2,i}(s) \right\}^2,$$

respectively

Design parameters

- Gap times($s_{k,i,j}$): $\mathrm{Poisson}(\lambda_k^s = \rho\lambda_k)$
  - Observation times: $t_{k,i,j} = t_{k,i,j-1} + s_{k,i,j}$ with $t_{k,i,0} \equiv 0$
- Two types of dependency
  - $\rho = 1$
  - $\rho \sim \mathrm{Gamma}(0.2, 5)$
- Two types of mean function
  - $v(t) = 0.1 + 0.8t$
  - $w(t) = 0.1 + 0.7t + 0.3t^{0.5}$
- Error term($\epsilon_{k,i}$): $\mathrm{N}(0, 0.7^2)$
- Survival times($D_{k,i,j}$): $\mathrm{Exp}(\lambda_k^D)$ on $(2, \infty)$, i.e., Generated from a two-parameter exponential distribution
- Censoring times($C_{k,i}$): Fixed censoring at 6.7

Setup for parameters

- For Group 1, $\lambda_1 = 1.0$ and $\lambda_1^D = 0.4$
- For size, $m_1(t) = v(t) = m_2(t)$, $\lambda_1 = \lambda_2$, and $\lambda_1^D = \lambda_2^D$
- For power, $m_1(t) = v(t)$ and $m_2(t) = w(t)$, i.e., $m_1(t) \neq m_2(t)$, $\lambda_1 \neq \lambda_2$, or $\lambda_1^D \neq \lambda_2^D$
- Sample sizes: $n_1 = n_2 = 50$ or 100
- Based on 1,000 samples

Size and powers when $\rho = 1$

		$m_1(t) = m_2(t)$		$m_1(t) \neq m_2(t)$	
λ_2	λ_2^D	$n = 50$	$n = 100$	$n = 50$	$n = 100$
1.0	0.4	0.050	0.052	0.190	0.278
1.0	0.2	0.080	0.096	0.264	0.386
1.0	0.1	0.114	0.196	0.312	0.438
0.6	0.4	0.992	1.000	1.000	1.000
0.8	0.4	0.510	0.744	0.850	0.982
0.9	0.4	0.144	0.230	0.536	0.750

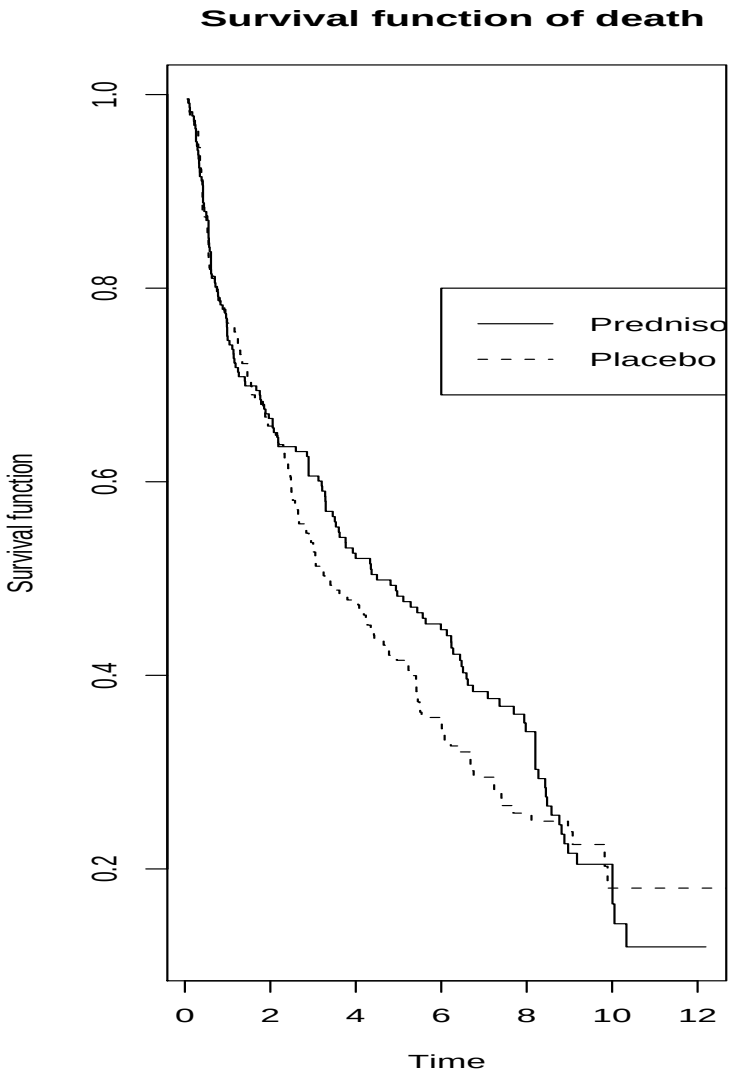
Size and powers when $\rho \sim \mathrm{Gamma}(0, 2, 5)$

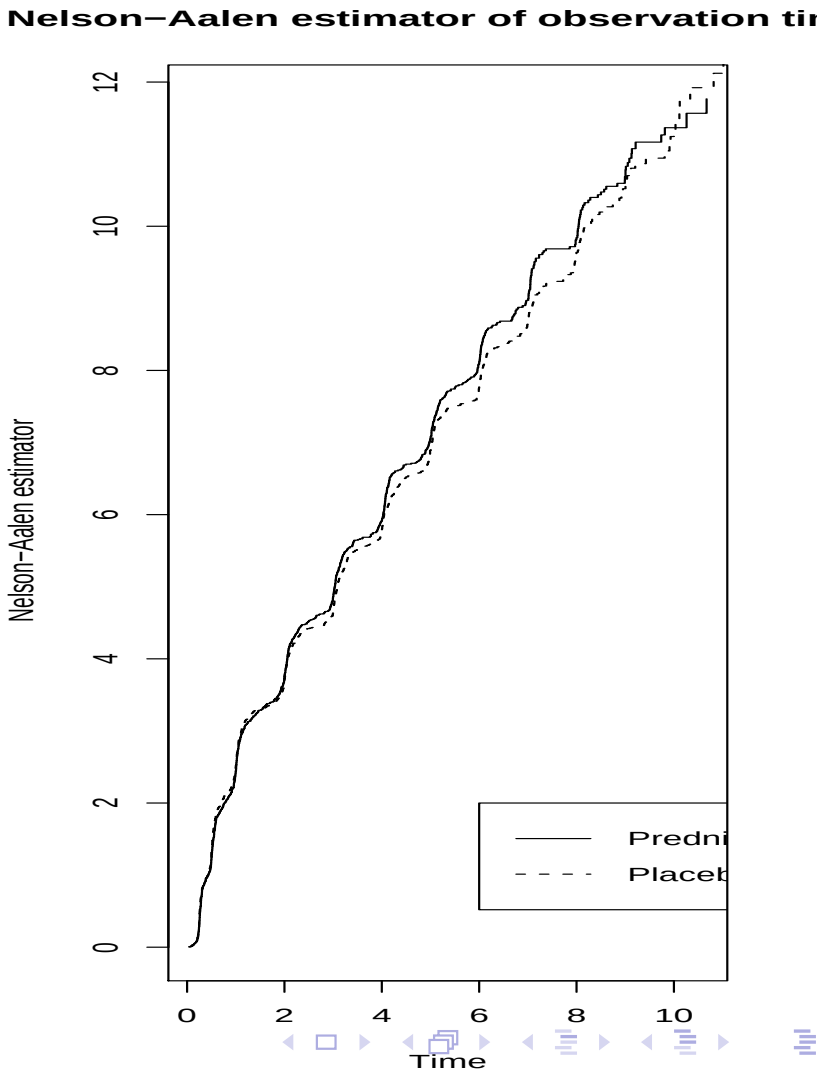
		$m_1 = m_2$		$m_1 \neq m_2$	
λ_2	λ_2^D	$n = 50$	$n = 100$	$n = 50$	$n = 100$
1.0	0.4	0.058	0.052	0.110	0.141
1.0	0.2	0.092	0.168	0.154	0.260
1.0	0.1	0.152	0.264	0.300	0.310
0.6	0.4	0.892	0.982	0.968	0.996
0.8	0.4	0.352	0.486	0.604	0.836
0.9	0.4	0.132	0.133	0.334	0.462

CSL 1 data set

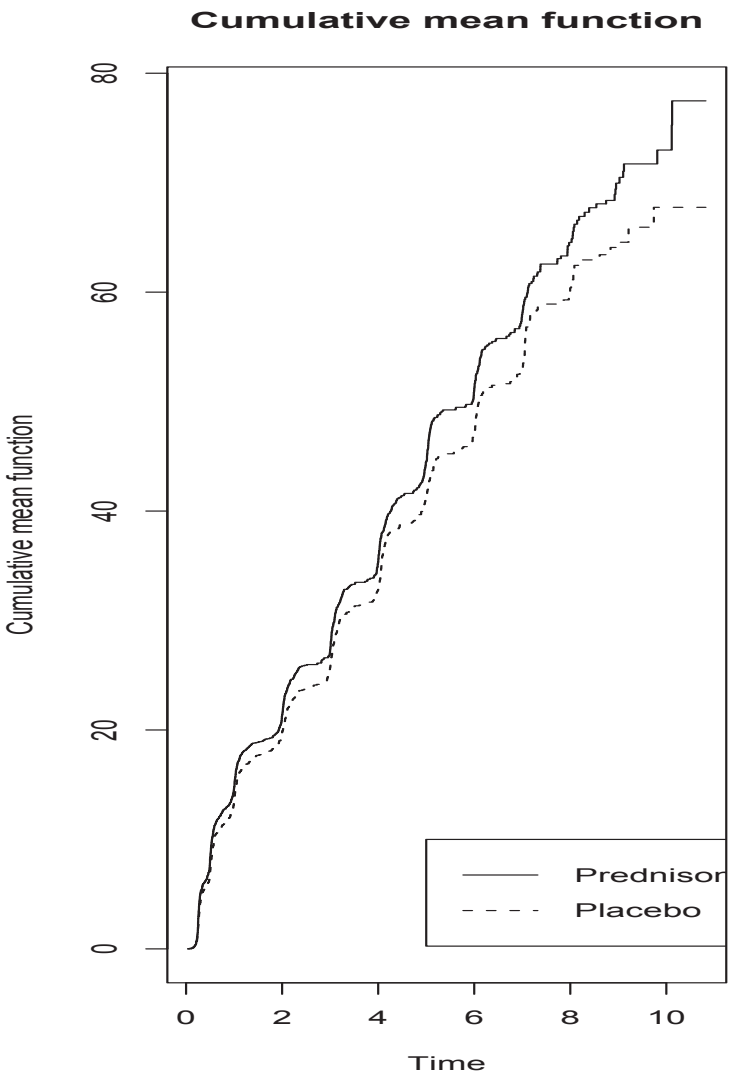
- Data set from 446 patients with liver cirrhosis conducted by the Copenhagen Study Group for Liver Diseases (Schlichting et al., Hepatology, 1983)
- Outcome: Prothrombin index, a measurement based on a blood test of coagulation factors II, VII, and X produced by the liver
- Placebo: 257 (165: died, 92: censored), Prednisone-treated: 189 (105: died, 84: censored)
- Scheduled visiting times: At the entry, three, six and twelve months after treatment and thereafter once a year, BUT the actual visiting times varied considerably around the scheduled times
- # of visits ranged from 1 to 16
- Average visit numbers: 5.45 (Placebo) and 5.72 (Prednisone-treated)
- $\tilde{Q}_{\mathrm{LR}} = \sqrt{n_1 n_2 / n} Q_{\mathrm{LR}} = -2.809$ ($p-$value=0.0005)

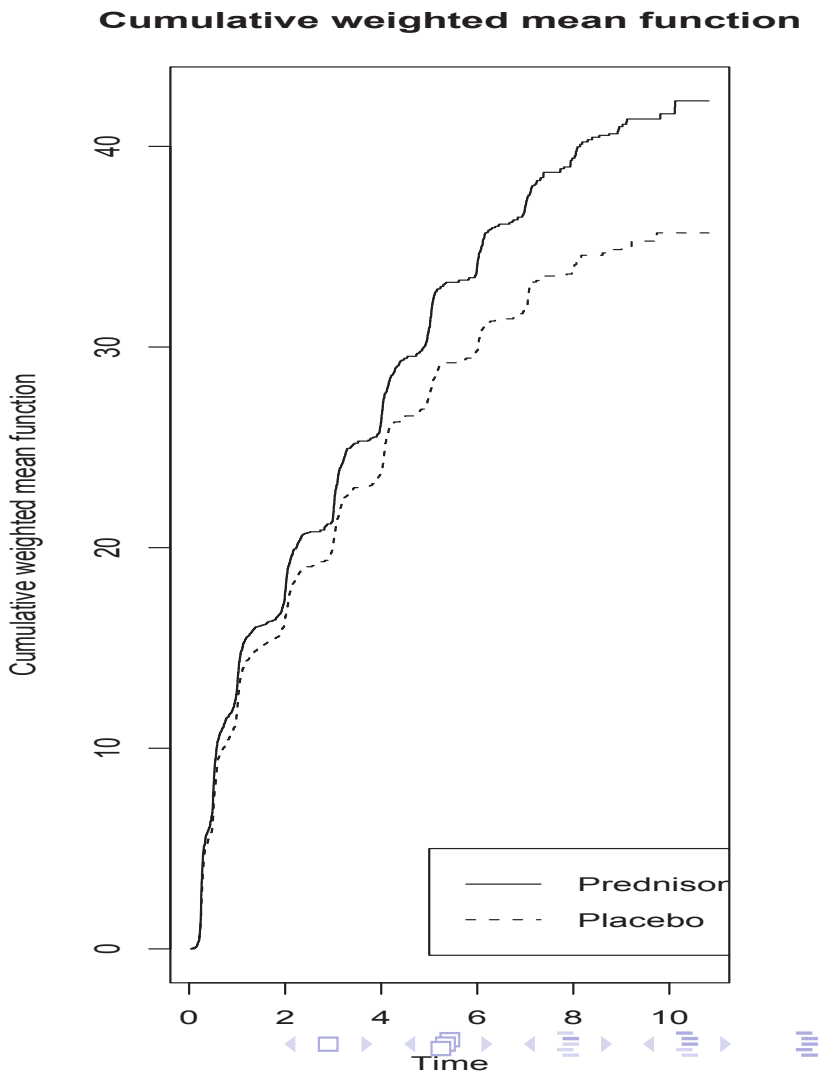
Plots of survival curves & Nelson-Aalen estimators of visiting times





Plots of cumulative mean functions & weighted cumulative mean functions





Conclusion & extension

- Propose test statistics to compare the mean functions of two groups for longitudinal data with group-specific observation and termination times
- According to simulations, it controls well a significance level and also its power seems to be reasonable for several combinations of the distribution of the observation times and the intensity of termination time
- Extend to the longitudinal regression data with covariate-specific observation and termination times

Thank You!