Quantiles of the conditional residual lifetime

ABRAMS, Steven; JANSSEN, Paul & VERAVERBEKE, Noel (2021) Quantiles of the conditional residual lifetime. In: Statistics (Berlin), 55 (6) , p. 1271-1290.

DOI: 10.1080/02331888.2021.2006661
Handle: http://hdl.handle.net/1942/36478

ORIGINAL ARTICLE

Quantiles of the conditional residual lifetime

S. Abrams[a,b] and P. Janssen[a,c] and N. Veraverbeke[a,c]

[a]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Data Science Institute (DSI), University of Hasselt, Diepenbeek, Belgium
[b]Global Health Institute (GHI), Family Medicine and Population Health (FAMPOP), University of Antwerp, Wilrijk, Belgium
[c]North-West University, Potchefstroom Campus, Potchefstroom, South Africa

ARTICLE HISTORY
Compiled November 3, 2021

ABSTRACT
The study of the residual lifetime received considerable attention in survival analysis and in other disciplines like reliability theory and actuarial science. The quantile residual lifetime function, the inverse of the residual lifetime distribution $P(T_1 - t_1 \leq y \mid T_1 > t_1)$, provides an interesting and well-studied measure to analyse residual lifetimes. In this paper we generalize the residual lifetime distribution and the quantile residual lifetime function by adding an extra conditioning of the form $\{T_2 \leq t_2\}$ or $\{T_2 > t_2\}$, where T_2 is a second variable containing extra information on T_1. We propose, for right-censored lifetimes, nonparametric estimators for this generalized conditional remaining lifetime distribution and the corresponding quantile function and we derive the asymptotic theory. In a simulation study, we show the good performance of the newly proposed quantile estimators and we discuss an application to real data on primary biliary cirrhosis.

KEYWORDS
Asymptotic representation; bivariate distribution; conditional residual lifetime; quantiles; right censoring

1. Introduction

Inference on the remaining or residual lifetime (T_1) of patients is of ultimate interest in many medical trials. Residual lifetimes are also important in other disciplines including reliability theory (remaining lifetime of devices) and actuarial science (remaining lifetime of policyholders). In clinical trials event times are typically subject to right censoring and, often, additional information (T_2), such as a prognostic variable or index, is available. Comparing the group of patients having $\{T_2 > t_2\}$ with the group of patients having $\{T_2 \leq t_2\}$ provides a fairly simple way to describe the impact of T_2 on T_1. In this paper we broaden the concept of the conditional residual lifetime distribution by adding an extra conditioning on either $\{T_2 > t_2\}$ or $\{T_2 \leq t_2\}$. Although the role of T_2, as covariate on which conditioning happens, is different from the one of T_1 (i.e., usually an event time of interest subject to censoring), hence, T_2 is not necessarily a time variable, we do use a similar notation for both variables to be in

CONTACT S. Abrams. Email: steven.abrams@uhasselt.be

line with relevant literature regarding the estimation of conditional residual lifetime distributions.

Functionals based on these new conditional residual lifetime distributions $P(T_1 \leq t_1 + y \mid T_1 > t_1, T_2 \leq t_2)$ and $P(T_1 \leq t_1 + y \mid T_1 > t_1,\ T_2 > t_2)$ can be used to quantify the impact of T_2 on T_1, e.g., the mean of the conditional residual lifetimes is a possible way to summarize remaining life expectancy data. However for censored data it is more appealing (parallel to the Kaplan-Meier case) to consider the median conditional residual lifetime. In fact we can consider other quantiles of the conditional residual lifetime.

In some recent papers the study of the (non)parametric estimation of quantiles of residual lifetimes conditioned on $\{T_2 = t_2\}$, i.e., quantiles corresponding to $P(T_1 \leq t_1 + y \mid T_1 > t_1,\ T_2 = t_2)$ for right-censored lifetime data received attention (often taking a regression perspective). See, e.g., [1] and [2]. But corresponding results for quantiles of residual lifetimes categorized according to the conditioning $\{T_2 > t_2\}$, resp. $\{T_2 \leq t_2\}$ are new.

In absence of additional information (T_2) the quantiles corresponding to the conditional distribution $P(T_1 \leq t_1 + y \mid T_1 > t_1)$ have been studied by many authors and the advantages of quantiles when compared to the mean residual lifetime $E(T_1 - t_1 \mid T_1 > t_1)$ have been discussed, see, e.g., Schmittlein and Morrison [3] and Joe and Proshan [4].

As main methodological contribution we establish the asymptotic normality results for nonparametric estimators for quantiles of the conditional residual lifetime for observed data of form $D_i = (Z_i, \delta_i, T_{2i})$ where $Z_i = T_{1i} \wedge C_i$ with C_i the censoring variable (which is allowed to depend on T_{2i}) and δ_i the censoring indicator, $i = 1, \ldots, n$ (see Section 7). Note that related work by Kayid et al. [5] does not allow that the lifetime is subject to right censoring.

On the practical side we show - in a small simulation study - the good finite sample performance of the proposed nonparametric estimators of quantiles of the conditional residual lifetime. In the simulation we use Clayton, Gumbel or Farlie-Gumbel-Morgenstern copulas to model the association between T_1 and T_2 (Section 8). Computational aspects (simulation algorithm) are given in Appendix A of the Supplement. We further demonstrate (Section 9) the use of the proposed quantile estimator for data on primary biliary cirrhosis (PBC) (see Fleming and Harrington [6]). A second application on advanced lung cancer data (a well-known data set derived from a study of the North Central Cancer Treatment Group) (see Therneau and Grambsch [7]) is discussed in the Supplementary Material. There we also show that the ratio of the quantiles of the residual lifetimes conditioned on $\{T_2 > t_2\}$ and $\{T_2 \leq t_2\}$ can be used for risk comparison between groups. This ratio gives visual information on the association between T_1 and T_2 and provides an alternative for risk assessment based on the ratio of conditional hazard rate functions, i.e., based on $\lambda(t_1 \mid T_2 > t_2)/\lambda(t_1 \mid T_2 \leq t_2)$. Nonparametric estimation of the hazard-based risk ratio has been studied in Abrams et al. [8].

Appropriate definitions and some working tools, including asymptotic i.i.d. representations for empirical versions of the conditional distribution functions $P(T_1 \leq t_1 + y \mid T_1 > t_1,\ T_2 \leq t_2)$ and $P(T_1 \leq t_1 + y \mid T_1 > t_1, T_2 > t_2)$, needed to prove the asymptotic normality of the nonparametric quantile estimators in Section 7 are collected in Sections 2–6. Some suggestions for future work are collected in Section 10.

2. Conditional residual lifetime

Let (T_1, T_2) be a random vector, $T_1 \geq 0$, $T_2 \geq 0$, with joint distribution $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$ and marginals $F_1(t_1) = P(T_1 \leq t_1)$, $F_2(t_2) = P(T_2 \leq t_2)$.

We define the conditional distribution function of T_1 given that $T_2 \leq t_2$:

$$\begin{aligned} F(t_1 \mid T_2 \leq t_2) &= P(T_1 \leq t_1 \mid T_2 \leq t_2) = \frac{F(t_1, t_2)}{F_2(t_2)} \\ &=: \widetilde{F}_{t_2}(t_1) \text{ (shorthand notation)}, \end{aligned} \tag{1}$$

and the conditional residual lifetime distribution function of T_1 at t_1, given that $T_2 \leq t_2$:

$$\begin{aligned} &P(T_1 - t_1 \leq y \mid T_1 > t_1, T_2 \leq t_2) = \frac{P(t_1 < T_1 \leq t_1 + y, T_2 \leq t_2)}{P(T_1 > t_1, T_2 \leq t_2)} \\ &= \frac{F(t_1 + y, t_2) - F(t_1, t_2)}{F_2(t_2) - F(t_1, t_2)} = \frac{\widetilde{F}_{t_2}(t_1 + y) - \widetilde{F}_{t_2}(t_1)}{1 - \widetilde{F}_{t_2}(t_1)}. \end{aligned} \tag{2}$$

For $0 < p < 1$, we define the p-th quantile of the conditional residual lifetime of T_1, given that $T_2 \leq t_2$ as

$$\begin{aligned} \widetilde{Q}(p \mid t_1, t_2) &= \inf \left\{ y : \frac{\widetilde{F}_{t_2}(t_1 + y) - \widetilde{F}_{t_2}(t_1)}{1 - \widetilde{F}_{t_2}(t_1)} \geq p \right\} \\ &= \inf \left\{ y : \widetilde{F}_{t_2}(t_1 + y) \geq \widetilde{F}_{t_2}(t_1) + p[1 - \widetilde{F}_{t_2}(t_1)] \right\} \\ &= -t_1 + \widetilde{F}_{t_2}^{-1}[p + (1 - p)\widetilde{F}_{t_2}(t_1)], \end{aligned} \tag{3}$$

where $\widetilde{F}_{t_2}^{-1}(p) = \inf\{y : \widetilde{F}(y) \geq p\}$ is the inverse of $\widetilde{F}_{t_2}$.
For conditioning on $T_2 > t_2$ we can consider the corresponding quantities:

$$\begin{aligned} F(t_1 \mid T_2 > t_2) &= P(T_1 \leq t_1 \mid T_2 > t_2) = \frac{F_1(t_1) - F(t_1, t_2)}{1 - F_2(t_2)} \\ &=: \widetilde{\widetilde{F}}_{t_2}(t_1) \text{ (shorthand notation)}, \end{aligned} \tag{4}$$

$$P(T_1 - t_1 \leq y \mid T_1 > t_1, T_2 > t_2) = \frac{\widetilde{\widetilde{F}}_{t_2}(t_1 + y) - \widetilde{\widetilde{F}}_{t_2}(t_1)}{1 - \widetilde{\widetilde{F}}_{t_2}(t_1)}, \tag{5}$$

$$\widetilde{\widetilde{Q}}(p \mid t_1, t_2) = -t_1 + \widetilde{\widetilde{F}}_{t_2}^{-1}[p + (1 - p)\widetilde{\widetilde{F}}_{t_2}(t_1)]. \tag{6}$$

Nonparametric estimation of $\widetilde{Q}(p \mid t_1, t_2)$ and $\widetilde{\widetilde{Q}}(p \mid t_1, t_2)$ will require nonparametric estimators for the conditional distribution functions $\widetilde{F}_{t_2}$ and $\widetilde{\widetilde{F}}_{t_2}$ and their quantile

functions $\widetilde{F}_{t_2}^{-1}$ and $\widetilde{\widetilde{F}}_{t_2}^{-1}$, respectively. From (1) and (4) we see that we need estimators for the bivariate distribution function $F(t_1, t_2)$ and the marginals $F_1(t_1)$ and $F_2(t_2)$.

3. Setting the scene

The setting for this paper is the following. For the bivariate random vector (T_1, T_2), the first component T_1 is subject to random right censoring by a censoring variable C. Let $Z = T_1 \wedge C$ and $\delta = I(T_1 \leq C)$. The second component T_2 is always observed (uncensored). The censoring variable C is allowed to depend on T_2 and it is assumed that, given T_2, the variables T_1 and C are independent.

The observed data are

$$D_i = (Z_i, \delta_i, T_{2i}), \qquad i = 1, \ldots, n$$

where $Z_i = T_{1i} \wedge C_i$ and the censoring indicator $\delta_i = I(T_{1i} \leq C_i)$. We construct estimators based on a sample $D_i \overset{iid}{\sim} D = (Z, \delta, T_2)$ for $i = 1, \ldots, n$.

We introduce some further notation, related to the observations:

$$\begin{aligned} H(z \mid t) &= P(Z \leq z \mid T_2 = t), \\ H^u(z \mid t) &= P(Z \leq z,\ \delta = 1 \mid T_2 = t). \end{aligned}$$

Because of conditional independence of T_1 and C, we have

$$1 - H(z \mid t) = [1 - F(z \mid t)][1 - G(z \mid t)],$$

where

$$F(z \mid t) = P(T_1 \leq z \mid T_2 = t) \text{ and } G(z \mid t) = P(C \leq z \mid T_2 = t).$$

We also have

$$H^u(z \mid t) = \int_0^z [1 - G(s- \mid t)] dF(s \mid t).$$

In estimation problems with right-censored observations, the support of the distribution of Z plays an important role. If there is no conditioning on T_2 (e.g., on $\{T_2 \leq t_2\}$), then we know that it is only possible to estimate $F_1(t_1)$ for t_1 values below the upper endpoint of support of the distribution of Z. Also known is that uniformity of the remainder term in the asymptotic representation can only be achieved if t_1 stays strictly away from this upper endpoint of support. In the presence of T_2, it will only be possible to estimate $F(t_1, t_2)$ for (t_1, t_2)-values for which t_1 is below the upper endpoint of the support of the conditional distribution of Z, given $T_2 = t$, for all $t \leq t_2$. Indeed, as worked out in the next section, the estimator for $F(t_1, t_2)$ will be obtained from relation (7) by plugging in the empirical distribution function for the (uncensored) observations of $F_2(t)$ and the conditional Kaplan-Meier estimator for $F(t_1|t)$. Therefore, we will have to stay strictly away from the right endpoint of support of F_2 as well as from the right endpoint of support of $P(Z \leq z|T_2 = t)$, for all $t \in [0, t_2]$ (the range of

the integral in (7)).
Hence, in order to define the domain for our estimators, we introduce the following notation (as in Akritas [9] and Akritas and Van Keilegom [10]):

$$\begin{aligned}\tau_1(t) &= \text{any number} < \inf\{z : H(z \mid t) = 1\}, \text{ the upper endpoint of the} \\ &\quad \text{support of } H(z \mid t) = P(Z \leq z \mid T_2 = t),\end{aligned}$$

$$\begin{aligned}\tau_2 &= \text{any number}, 0 \leq \tau_2 < \inf\{t : F_2(t) = 1\}, \text{ the upper endpoint of the} \\ &\quad \text{support of } F_2(t) = P(T_2 \leq t).\end{aligned}$$

Throughout, we will use the following domain:

$$\Omega = \left\{(t_1, t_2) : t_2 \leq \tau_2, t_1 \leq \inf_{t \leq t_2} \tau_1(t)\right\}.$$

4. Estimation of the joint distribution

The estimator for $F(t_1, t_2)$ is obtained from the relation

$$F(t_1, t_2) = \int_0^{t_2} F(t_1 \mid t) dF_2(t) \tag{7}$$

by plugging in estimators $F_n(t_1 \mid t)$ for $F(t_1 \mid t)$ and $F_{2n}(t)$ for $F_2(t)$ where $F_{2n}(t)$ represents the usual empirical distribution function as estimator for $F_2(t)$. This gives

$$F_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} F_n(t_1 \mid T_{2i}) I(T_{2i} \leq t_2). \tag{8}$$

Here we describe some properties of this estimator for $F(t_1, t_2)$, which appear in papers by Akritas [9] and Akritas and Van Keilegom [10].

An estimator for $F(t_1 \mid t)$ is given by the Beran estimator [11,12], which is a generalization of the Kaplan-Meier estimator and therefore is sometimes referred to as the conditional Kaplan-Meier estimator. For any $(t_1, t) \in \Omega$ it is defined as

$$\begin{aligned}F_n(t_1 \mid t) &= 1 - \prod_{\substack{i=1 \\ Z_i \leq t_1 \\ \delta_i = 1}}^{n} \left(1 - \frac{w_{ni}(t, h_n)}{\sum_{j=1}^{n} w_{nj}(t, h_n) I(Z_j \geq Z_i)}\right) \\ &= 1 - \prod_{Z_{(i)} \leq t_1}^{n} \left(1 - \frac{w_{n(i)}(t, h_n)}{1 - \sum_{j=1}^{i-1} w_{n(j)}(t, h_n)}\right)^{\delta_{(i)}}\end{aligned}$$

where $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$ denote the ordered Z_j-values, $j = 1, \ldots, n$, and $\delta_{(j)}$ represents the censoring indicator for $Z_{(j)}$. The weights $w_{ni}(t, h_n)$ (and $w_{n(i)}(t, h_n)$

corresponding to $Z_{(j)}$) are Nadaraya-Watson weights with

$$w_{ni}(t, h_n) = K\left(\frac{t - T_{2i}}{h_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{t - T_{2j}}{h_n}\right)$$

where K is a known probability density function (kernel) and $\{h_n\}$ a sequence of positive constants, tending to 0 as $n \to \infty$ (bandwidth sequence). Note that in [9], nearest neighbour weights were used instead of the aforementioned ones. Alternatively, Gasser-Müller weights could be considered. Note that putting the weights $w_{ni}(t, h_n)$ (and $w_{n(i)}(t, h_n)$) all equal to n^{-1} leads to the classical Kaplan-Meier estimator for the cumulative distribution function.

As Lo and Singh [13] did for the ordinary Kaplan-Meier estimator, an almost sure asymptotic representation has been proved for the Beran estimator in Van Keilegom and Veraverbeke [12]. This result is used to propose the following approximation

$$F_n(t_1 \mid T_{2i}) - F(t_1 \mid T_{2i}) \cong \sum_{j=1}^{n} w_{nj}(T_{2i}, h_n)\xi(t_1, Z_j, \delta_j, T_{2i})$$

where the function $\xi(t_1, Z, \delta, t)$ is given by

$$\begin{aligned}
&\xi(t_1, Z, \delta, t) \\
&= [1 - F(t_1 \mid t)] \left\{ - \int_0^{Z \wedge t_1} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2} + \frac{I(Z \le t_1, \delta = 1)}{1 - H(Z \mid t)} \right\} \\
&= [1 - F(t_1 \mid t)] \left\{ \int_0^{t_1} \frac{I(Z \le s) - H(s \mid t)}{[1 - H(s \mid t)]^2} dH^u(s \mid t) \right. \\
&\left. + \frac{I(Z \le t_1, \delta_i = 1) - H^u(t_1 \mid t)}{1 - H(t_1 \mid t)} - \int_0^{t_1} \frac{I(Z \le s, \delta = 1) - H^u(s \mid t)}{[1 - H(s \mid t)]^2} dH(s \mid t) \right\}.
\end{aligned}$$

The conditional expectation of $\xi(t_1, Z, \delta, T_2)$ given $T_2 = t$ is equal to 0 for all t_1 and the conditional covariance of $\xi(t_1, Z, \delta, T_2)$ and $\xi(t_1', Z, \delta, T_2)$, given $T_2 = t$ is equal to

$$[1 - F(t_1 \mid t)][1 - F(t_1' \mid t)] \left\{ \int_0^{t_1 \wedge t_1'} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2} \right\},$$

[see, e.g., 12–14].

We make the following decomposition of (8):

$$\begin{aligned}
& F_n(t_1, t_2) - F(t_1, t_2) \\
= \; & \frac{1}{n} \sum_{i=1}^{n} \{F(t_1 \mid T_{2i}) I(T_{2i} \le t_2) - F(t_1, t_2)\} \\
& + \frac{1}{n} \sum_{i=1}^{n} \{F_n(t_1 \mid T_{2i}) - F(t_1 \mid T_{2i})\} I(T_{2i} \le t_2).
\end{aligned}$$

We now formulate the assumptions as they appear in Akritas and Van Keilegom [10].

(*A*1) $\dfrac{\log n}{nh_n} \to 0$, $nh_n^4 \to 0$;
K is a probability density function with support $[-1,1]$, K is twice differentiable, $\int uK(u)du = 0$.

(*A*2) $F_2(t_2)$ is three times continuously differentiable w.r.t. t_2;
$H(z \mid t_2)$ and $H^u(z \mid t_2)$ are twice continuously differentiable w.r.t. z and t_2 and for $(z, t_2) \in \Omega$, all derivatives are uniformly bounded.

As in Lemma's 3.2–3.4 of Akritas [9] it follows that, uniform in Ω,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} w_{nj}(T_{2i}, h_n)\xi(t_1, Z_j, \delta_j, T_2)\right] I(T_{2i} \le t_2) \\ = \; & \frac{1}{n}\sum_{i=1}^{n} \xi(t_1, Z_i, \delta_i, T_{2i}) I(T_{2i} \le t_2) + o_P(n^{-1/2}). \end{aligned}$$

Theorem 4.1 (Akritas [9], Akritas and Van Keilegom [10])**.** *Under assumptions* (*A*1) *and* (*A*2)*, we have for fixed* t_1 *and* t_2

$$F_n(t_1, t_2) = F(t_1, t_2) + \frac{1}{n}\sum_{i=1}^{n} \psi(t_1, Z_i, \delta_i, T_{2i}) + r_n(t_1, t_2)$$

with

$$\psi(t_1, Z_i, \delta_i, T_{2i}) = [F(t_1 \mid T_{2i}) I(T_{2i} \le t_2) - F(t_1, t_2)] + \xi(t_1, Z_i, \delta_i, T_{2i}) I(T_{2i} \le t_2)$$

and

$$\sup_{(t_1,t_2)\in\Omega} \mid r_n(t_1, t_2) \mid = o_P(n^{-1/2}).$$

Also, as $n \to \infty$,

$$n^{1/2}[F_n(t_1, t_2) - F(t_1, t_2)] \xrightarrow{d} N(0; \sigma^2(t_1, t_2))$$

where

$$\begin{aligned} \sigma^2(t_1, t_2) = & \int_0^{t_2} F^2(t_1|t) dF_2(t) - F^2(t_1, t_2) \\ & + \int_0^{t_2} [1 - F(t_1 \mid t)]^2 \left\{ \int_0^{t_1} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2} \right\} dF_2(t). \end{aligned}$$

Remark 1. The two terms in the expression for ψ are uncorrelated.

Remark 2. Beran estimators $F_n(t_1 \mid t)$ are known to have typical convergence rate $(nh_n)^{-1/2}$, but due to the averaging over the covariates, it turns out that the resulting estimator $F_n(t_1, t_2)$ has the faster rate of convergence $n^{-1/2}$.

5. Estimation of the conditional distribution functions of T_1 given $T_2 \leq t_2$ or given $T_2 > t_2$

For the conditional distribution function $\widetilde{F}_{t_2}(t_1)$ of T_1, given that $T_2 \leq t_2$, as in (3) we propose the following estimator

$$\widetilde{F}_{t_2,n}(t_1) = \frac{F_n(t_1,t_2)}{F_{2n}(t_2)}$$

with $F_n(t_1,t_2)$ as in (8) and $F_{2n}(t_2)$ the empirical distribution function of $F_2(t_2)$.

Theorem 5.1. *Under assumptions* $(A1)$ *and* $(A2)$*, we have for fixed* t_1 *and* t_2

$$\widetilde{F}_{t_2,n}(t_1) = \widetilde{F}_{t_2}(t_1) + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\psi}(t_1, Z_i, \delta_i, T_{2i}) + \widetilde{r}_n(t_1,t_2)$$

with

$$\widetilde{\psi}(t_1, Z_i, \delta_i, T_{2i}) = \frac{1}{F_2(t_2)}[F(t_1 \mid T_{2i})I(T_{2i} \leq t_2) - F(t_1,t_2)]$$
$$-\frac{F(t_1,t_2)}{F_2^2(t_2)}[I(T_{2i} \leq t_2) - F_2(t_2)] + \frac{1}{F_2(t_2)}\xi(t_1, Z_i, \delta_i, T_{2i})I(T_{2i} \leq t_2)$$

and

$$\sup_{(t_1,t_2)\in\Omega} \mid \widetilde{r}_n(t_1,t_2) \mid = o_P(n^{-1/2}).$$

The proof follows from Theorem 4.1 and Slutski's theorem applied to the linearization

$$\begin{aligned}\widetilde{F}_{t_2,n}(t_1) - \widetilde{F}_{t_2}(t_1) &= \frac{1}{F_{2n}(t_2)}[F_n(t_1,t_2) - F(t_1,t_2)] \\ &- \frac{F(t_1,t_2)}{F_{2n}(t_2)F_2(t_2)}[F_{2n}(t_2) - F_2(t_2)].\end{aligned}$$

A long but straightforward calculation gives the following expression for the covariance function:

$$\begin{aligned}&E[\widetilde{\psi}(t_1, Z, \delta, T_2)\widetilde{\psi}(t_1', Z, \delta, T_2)] \\ = \;& \frac{1}{F_2^2(t_2)}\int_0^{t_2} F(t_1 \mid t)F(t_1' \mid t)dF_2(t) - \frac{F(t_1,t_2)F(t_1',t_2)}{F_2^3(t_2)} \\ &+\frac{1}{F_2^2(t_2)}\int_0^{t_2}[1 - F(t_1 \mid t)][1 - F(t_1' \mid t)]\int_0^{t_1\wedge t_1'}\frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2}dF_2(t).\end{aligned} \tag{9}$$

For the conditional distribution function $\widetilde{\widetilde{F}}_{t_2}(t_1)$ of T_1, given $T_2 > t_2$, as in (4), we

propose the following estimator

$$\tilde{\tilde{F}}_{t_2,n}(t_1) = \frac{F_n(t_1,+\infty) - F_n(t_1,t_2)}{1 - F_{2n}(t_2)}.$$

Note that

$$F_n(t_1,+\infty) = \int_0^\infty F_n(t_1 \mid t) dF_{2n}(t) = \frac{1}{n}\sum_{i=1}^n F_n(t_1 \mid T_{2i})$$

where $F_n(t_1 \mid t)$ is the Beran estimator for $F(t_1 \mid t)$.

Again, by linearization and Theorem 4.1, we obtain

Theorem 5.2. *Under assumptions* $(A1)$ *and* $(A2)$*, we have for fixed* t_1 *and* t_2

$$\tilde{\tilde{F}}_{t_2,n}(t_1) = \tilde{\tilde{F}}_{t_2}(t_1) + \frac{1}{n}\sum_{i=1}^n \tilde{\tilde{\psi}}(t_1; Z_i, \delta_i, T_{2i}) + \tilde{\tilde{r}}_n(t_1,t_2)$$

with

$$\begin{aligned}
&\tilde{\tilde{\psi}}(t_1, Z_i, \delta_i, T_{2i}) = \\
&\frac{1}{1-F_2(t_2)}\left[F(t_1 \mid T_{2i}) - F_1(t_1)\right] + \frac{F_1(t_1) - F(t_1,t_2)}{(1-F_2(t_2))^2}\left[I(T_{2i} \le t_2) - F_2(t_2)\right] \\
&-\frac{1}{1-F_2(t_2)}\left[F_1(t_1 \mid T_{2i})I(T_{2i} \le t_2) - F(t_1,t_2)\right] \\
&+\frac{1}{1-F_2(t_2)}\xi(t_1, Z_i, \delta_i, T_{2i})I(T_{2i} > t_2)
\end{aligned}$$

and

$$\sup_{(t_1,t_2)\in\Omega} \mid \tilde{\tilde{r}}_n(t_1,t_2) \mid = o_P(n^{-1/2}).$$

Again, by direct calculation,

$$\begin{aligned}
&E[\tilde{\tilde{\psi}}(t_1, Z, \delta, T_2)\tilde{\tilde{\psi}}(t_1', Z, \delta, T_2)] \\
&= \frac{1}{[1-F_2(t_2)]^2}\int_{t_2}^\infty F(t_1 \mid t)F(t_1' \mid t) dF_2(t) \\
&-\frac{1}{[1-F_2(t_2)]^3}[F_1(t_1) - F(t_1,t_2)][F_1(t_1') - F(t_1',t_2)] \\
&+\frac{1}{[1-F_2(t_2)]^2}\int_{t_2}^\infty [1-F(t_1 \mid t)][1-F(t_1' \mid t)] \int_0^{t_1\wedge t_1'} \frac{dH^u(s \mid t)}{[1-H(s \mid t)]^2} dF_2(t).
\end{aligned} \tag{10}$$

6. Asymptotic representation for the quantile functions of $\widetilde{F}$ and $\widetilde{\widetilde{F}}$

In this section we deal with the quantile functions $\widetilde{F}_{t_2}^{-1}(p)$ and $\widetilde{\widetilde{F}}_{t_2}^{-1}(p)$ of the conditional distribution functions $\widetilde{F}_{t_2}(t_1)$ and $\widetilde{\widetilde{F}}_{t_2}(t_1)$ of T_1, given $T_2 \leq t_2$ and $T_2 > t_2$ respectively.

Asymptotic representations of quantile functions as sums of i.i.d. random variables can be obtained via so called Bahadur type theorems. We use here the version of Ghosh [15], which aims at a remainder term of order $o_P(n^{-1/2})$. Proofs are not given since they parallel that of a similar result in Gijbels and Veraverbeke [16, Theorem 2.1] for Kaplan-Meier quantiles.

For further use, we allow the order of the quantile to be random.

Theorem 6.1. *Assume* $(A1)$, $(A2)$ *and also*

$(\widetilde{A3})$ $0 < p < 1$ *and* $\widetilde{F}_{t_2}^{-1}(p) < \inf_{t \leq t_2} \tau_1(t)$.

$(\widetilde{A4})$ $F^{(1)}(t_1, t_2) = \dfrac{\partial}{\partial t_1} F(t_1, t_2)$ *exists at* $\left(\widetilde{F}_{t_2}^{-1}(p), t_2\right)$ *and* $F^{(1)}\left[\widetilde{F}_{t_2}^{-1}(p), t_2\right] > 0$;

$F(t_1 \mid t_2)$ *is Lipschitz continuous in* t_1 *for* t_1 *in a neighborhood of* $\widetilde{F}_{t_2}^{-1}(p)$.

If $\{p_n\}$ *is a sequence of random variables with* $p_n - p = O_P(n^{-1/2})$, *then, as* $n \to \infty$,

$$\widetilde{F}_{t_2,n}^{-1}(p_n) = \widetilde{F}_{t_2}^{-1}(p) + \frac{p_n - \widetilde{F}_{t_2,n}\left[\widetilde{F}_{t_2}^{-1}(p)\right]}{\widetilde{f}_{t_2}\left[\widetilde{F}_{t_2}^{-1}(p)\right]} + o_P(n^{-1/2})$$

where

$$\widetilde{f}_{t_2}(t_1) = \frac{\partial}{\partial t_1}\widetilde{F}_{t_2}(t_1) = \frac{F^{(1)}(t_1, t_2)}{F_2(t_2)}.$$

Theorem 6.2. *Assume* $(A1)$, $(A2)$ *and also*

$(\widetilde{\widetilde{A3}})$ $0 < p < 1$ *and* $\widetilde{\widetilde{F}}_{t_2}^{-1}(p) < \inf_{t \leq t_2} \tau_1(t)$.

$(\widetilde{\widetilde{A4}})$ $F^{(1)}(t_1, t_2) = \dfrac{\partial}{\partial t_1} F(t_1, t_2)$ *exists at* $\left(\widetilde{\widetilde{F}}_{t_2}^{-1}(p), t_2\right)$ *and* $F^{(1)}\left[\widetilde{\widetilde{F}}_{t_2}^{-1}(p), t_2\right] > 0$;

$f_1(t_1) = F_1'(t_1)$ *exists at* $\widetilde{\widetilde{F}}_{t_2}^{-1}(p)$;

$F(t_1 \mid t_2)$ *is Lipschitz continuous in* t_1 *for* t_1 *in a neighborhood of* $\widetilde{\widetilde{F}}_{t_2}^{-1}(p)$.

If $\{p_n\}$ *is a sequence of random variables with* $p_n - p = O_P(n^{-1/2})$, *then, as* $n \to \infty$,

$$\widetilde{\widetilde{F}}_{t_2,n}^{-1}(p) = \widetilde{\widetilde{F}}_{t_2}^{-1}(p) + \frac{p_n - \widetilde{\widetilde{F}}_{t_2,n}\left[\widetilde{\widetilde{F}}_{t_2}^{-1}(p)\right]}{\widetilde{\widetilde{f}}_{t_2}\left[\widetilde{\widetilde{F}}_{t_2}^{-1}(p)\right]} + o_P(n^{-1/2})$$

where

$$\widetilde{\widetilde{f}}_{t_2}(t_1) = \frac{\partial}{\partial t_1}\widetilde{\widetilde{F}}_{t_2}(t_1) = \frac{f_1(t_1) - F^{(1)}(t_1, t_2)}{1 - F_2(t_2)}.$$

7. Asymptotic normality for the quantiles of the conditional residual lifetime

Natural plug-in estimators for $\widetilde{Q}(p \mid t_1, t_2)$ in (3) and $\widetilde{\widetilde{Q}}(p \mid t_1, t_2)$ in (6) are

$$\widetilde{Q}_n(p \mid t_1, t_2) = -t_1 + \widetilde{F}^{-1}_{t_2,n}[p + (1-p)\widetilde{F}_{t_2,n}(t_1)]$$

and

$$\widetilde{\widetilde{Q}}_n(p \mid t_1, t_2,) = -t_1 + \widetilde{\widetilde{F}}^{-1}_{t_2,n}[p + (1-p)\widetilde{\widetilde{F}}_{t_2,n}(t_1)]$$

where $\widetilde{F}_{t_2,n}$ and $\widetilde{\widetilde{F}}_{t_2,n}$ are the estimators discussed in Section 5.
Denote $\widetilde{q} = p+(1-p)\widetilde{F}_{t_2}(t_1)$ and $\widetilde{\widetilde{q}} = p+(1-p)\widetilde{\widetilde{F}}_{t_2}(t_1)$ and let $\widetilde{q}_n = p+(1-p)\widetilde{F}_{t_2,n}(t_1)$ and $\widetilde{\widetilde{q}}_n = p + (1-p)\widetilde{\widetilde{F}}_{t_2,n}(t_1)$ be the empirical counterparts.
We have the following asymptotic normality results, where

$$V(p, v, t_1, t) = (1-p)^2[1 - F(t_1 \mid t)]^2 \int_0^{t_1} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2}$$
$$+[1 - F(v \mid t)]^2 \int_0^{v} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2}$$
$$-2(1-p)[1 - F(t_1 \mid t)(1 - F(v \mid t)] \int_0^{t_1 \wedge v} \frac{dH^u(s \mid t)}{[1 - H(s \mid t)]^2}.$$

Theorem 7.1. *Assume* $(A1)$, $(A2)$ *and also* $(\widetilde{A3})$, $(\widetilde{A4})$ *with* p *replaced by* $\widetilde{q}$. *Then, as* $n \to \infty$,

$$n^{1/2}[\widetilde{Q}_n(p \mid t_1, t_2) - \widetilde{Q}(p \mid t_1, t_2)] \xrightarrow{d} N(0; \widetilde{\sigma}^2_p(t_1, t_2))$$

where

$$\widetilde{\sigma}^2_p(t_1, t_2) = \frac{1}{\widetilde{f}^2_{t_2}[\widetilde{F}^{-1}_{t_2}(q)]} \left\{ \frac{1}{F_2^2(t_2)} \int_0^{t_2} V(p, \widetilde{F}^{-1}_{t_2}(\widetilde{q}), t_1, t) dF_2(t) \right.$$
$$+\frac{1}{F_2^2(t_2)} \int_0^{t_2} [(1-p)F(t_1 \mid t) - F(\widetilde{F}_{t_2}{}^{-1}(\widetilde{q}) \mid t)]^2 dF_2(t)$$
$$\left. -\frac{1}{F_2^3(t_2)}[(1-p)F(t_1, t_2) - F(\widetilde{F}^{-1}_{t_2}(\widetilde{q}), t_2)]^2 \right\}.$$

Proof. By using Theorem 5.1 and Theorem 6.1:

$$\widetilde{Q}_n(p \mid t_1, t_2) - \widetilde{Q}(p \mid t_1, t_2) = \frac{1}{\widetilde{f}_{t_2}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})]} \left\{ \widetilde{q}_n - \widetilde{F}_{t_2,n}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})] \right\} + o_P(n^{-1/2})$$
$$= \frac{1}{\widetilde{f}_{t_2}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})]} \left(\widetilde{q}_n - \widetilde{q} - \left\{ \widetilde{F}_{t_2,n}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})] - \widetilde{F}_{t_2}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})] \right\} \right) + o_P(n^{-1/2})$$
$$= \frac{1}{\widetilde{f}_{t_2}[\widetilde{F}^{-1}_{t_2}(\widetilde{q})]} \frac{1}{n} \sum_{i=1}^{n} \left\{ (1-p)\widetilde{\psi}(t_1, Z_i, \delta_i, T_{2i}) - \widetilde{\psi}[\widetilde{F}^{-1}_{t_2}(\widetilde{q}), Z_i, \delta_i, T_{2i}] \right\} + o_P(n^{-1/2}).$$

Asymptotic normality follows from this representation. The formula for the asymptotic variance $\widetilde{\sigma}_p^2(t_1,t_2)$ is obtained by direct calculation and using the covariance expression (9) after Theorem 5.1:

$$\widetilde{\sigma}_p^2(t_1,t_2) = \frac{1}{\widetilde{f}_{t_2}^2[\widetilde{F}_{t_2}^{-1}(\widetilde{q})]} E\left\{\left[(1-p)\widetilde{\psi}(t_1,Z,\delta,T_2) - \widetilde{\psi}(\widetilde{F}_{t_2}^{-1}(\widetilde{q}),Z,\delta,T_2)\right]^2\right\}.$$

□

Remark 3. In case of no censoring, this asymptotic variance simplifies to

$$\begin{aligned}\widetilde{\sigma}_p^2(t_1,t_2) &= \frac{1}{\widetilde{f}_{t_2}^2[\widetilde{F}_{t_2}^{-1}(\widetilde{q})]} \times \\ &\left\{\frac{1}{F_2^2(t_2)}[(1-p)^2F_1(t_1) - F_1[\widetilde{F}_{t_2}^{-1}(\widetilde{q})] - 2(1-p)F_1(t_1)]\right. \\ &\left. - \frac{1}{F_2^3(t_2)}[(1-p)F(t_1,t_2) - F(\widetilde{F}_{t_2}^{-1}(\widetilde{q}),t_2)]^2\right\}.\end{aligned}$$

Remark 4. A further special case is the absence of T_2. If we let $t_2 \to \infty$, then $F_2(t_2) \to 1$, $F(t_1,t_2) \to F_1(t_1)$ and $\widetilde{F}_{t_2} = F_1$. The asymptotic variance becomes

$$\frac{p(1-p)[1-F_1(t_1)]}{f_1^2\{F_1^{-1}[p+(1-p)F_1(t)]\}} \tag{11}$$

which is a known result for the p-th quantiles of the residual lifetime in that case [16].

Theorem 7.2 gives the parallel asymptotic normality result for $\widetilde{\widetilde{Q}}_n(p \mid t_1,t_2)$. The proof is similar to Theorem 7.1 but now using Theorems 5.2 and 6.2 and the covariance expression (10) after Theorem 5.2.

Theorem 7.2.

Assume $(A1)$, $(A2)$ and also $(\widetilde{\widetilde{A3}})$, $(\widetilde{\widetilde{A4}})$ with p replaced by $\widetilde{\widetilde{q}}$.
Then, as $n \to \infty$,

$$n^{1/2}[\widetilde{\widetilde{Q}}_n(p \mid t_1,t_2) - \widetilde{\widetilde{Q}}(p \mid t_1,t_2)] \xrightarrow{d} N(0;\widetilde{\widetilde{\sigma}}_p^2(t_1,t_2))$$

where

$$\begin{aligned}\widetilde{\widetilde{\sigma}}_p^2(t_1,t_2) &= \frac{1}{\widetilde{\widetilde{f}}_{t_2}^2[\widetilde{\widetilde{F}}_{t_2}^{-1}(\widetilde{\widetilde{q}})]} \left\{\frac{1}{[1-F_2(t_2)]^2}\int_{t_2}^{\infty} V(p,\widetilde{\widetilde{F}}_{t_2}^{-1}(\widetilde{\widetilde{q}}),t_1,t)dF_2(t)\right. \\ &+\frac{1}{[1-F_2(t_2)]^2}\int_{t_2}^{\infty}[(1-p)F(t_1 \mid t) - F(\widetilde{\widetilde{F}}_{t_2}^{-1}(\widetilde{\widetilde{q}}) \mid t)]^2 dF_2(t) \\ &\left. - \frac{1}{[1-F_2(t_2)]^3}\left[(1-p)[F_1(t_1) - F(t_1,t_2)] - \left\{F_1(\widetilde{\widetilde{F}}_{t_2}^{-1}[\widetilde{\widetilde{q}}]) - F[\widetilde{\widetilde{F}}_{t_2}^{-1}(\widetilde{\widetilde{q}}),t_2]\right\}\right]\right\}.\end{aligned}$$

Remark 5. We can again simplify in case of no censoring. And if we put $t_2 = 0$,

$F_2(t_2) = 0$, $F(t_1, t_2) = 0$, $\widetilde{\widetilde{F}}_{t_2} = F_1$, we again obtain expression (11) as in Gijbels and Veraverbeke [16].

8. Simulation study

To evaluate the finite sample behaviour of the proposed quantile estimators, we set up a simulation study.

8.1. Simulation set-up

For T_j, $j = 1, 2$, we use a Weibull distribution with $s_j = 1.5$ as shape parameter and $d_j = 0.5$ as decay parameter (i.e., the scale parameter is d_j^{-1/s_j}), i.e.,

$$F_j(t_j) = 1 - \exp\left(-d_j t_j^{s_j}\right), \quad j = 1, 2. \tag{12}$$

For the censoring random variable C we use a Weibull distribution with shape parameter $s_C = 1.5$, for the decay parameter we consider two cases: $d_C = 0.15$ and $d_C = 0.85$. Since $P(C < T_1) = d_C/(d_C + d_1)$, we have approximately 23% (moderate) censoring for $d_C = 0.15$ and approximately 63% (heavy) censoring for $d_C = 0.85$.
In order to arrive at bivariate Weibull distributions, we use

1. the Clayton copula, $\theta \geq 0$,

$$C_C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta};$$

2. the Gumbel copula, $\theta \geq 1$,

$$C_G(u, v) = \exp\left(-\left\{[-\log(u)]^\theta + [-\log(v)]^\theta\right\}^{1/\theta}\right);$$

3. the Farlie-Gumbel-Morgenstern (FGM) copula, $0 \leq \theta \leq 1$,

$$C_{FGM}(u, v) = uv\left[1 + \theta(1 - u)(1 - v)\right].$$

Let $F(t_1, t_2) = C\left[F_1(t_1), F_2(t_2)\right]$, with $C \in \mathcal{C}_\theta$ (i.e., a one-parameter copula family). To obtain $\widetilde{Q}(p \mid t_1, t_2)$ we need to solve the following equation with respect to y

$$C(u(y), v) = pv + (1 - p)C(u, v), \tag{13}$$

where $u(y) = F_1(t_1 + y)$, $u = u(0) = F_1(t_1)$, $v = F_2(t_2)$. We use, as shorthand notation,

$$B(p \mid t_1, t_2, \theta) = pv + (1 - p)C(u, v).$$

For the Clayton, resp. the Gumbel, copula we have an explicit expression for $\widetilde{Q}(p \mid t_1, t_2)$. We can, indeed, show that for the Clayton copula (for arbitrary marginals F_1 and F_2)

$$\widetilde{Q}(p \mid t_1, t_2) = -t_1 + F_1^{-1}\left[A^{-1/\theta}(p \mid t_1, t_2, \theta)\right] \tag{14}$$

with

$$A(p \mid t_1, t_2, \theta) = 1 - F_2^{-\theta}(t_2) + [B(p \mid t_1, t_2, \theta)]^{-\theta}$$
$$= 1 - F_2^{-\theta}(t_2) + \left[pF_2(t_2) + (1-p)(F_1^{-\theta}(t_1) + F_2^{-\theta}(t_2) - 1)^{-1/\theta}\right]^{-\theta}.$$

For the Gumbel copula, we have

$$\widetilde{Q}(p \mid t_1, t_2) = -t_1 + F_1^{-1}\left\{\exp\left[-\left(\{-\log[B(p \mid t_1, t_2, \theta)]\}^\theta - \{-\log[F_2(t_2)]\}^\theta\right)^{1/\theta}\right]\right\},$$
$$B(p \mid t_1, t_2, \theta) = pF_2(t_2) + (1-p)\exp\left[-\left(\{-\log[F_1(t_1)]\}^\theta + \{-\log[F_2(t_2)]\}^\theta\right)^{1/\theta}\right]. \quad (15)$$

For the FGM copula it can be shown that, for arbitrary marginals F_1 and F_2,

$$\widetilde{Q}(p \mid t_1, t_2) = -t_1 + F_1^{-1}(Y_0), \quad (16)$$

with Y_0 the smallest root of the quadratic equation

$$\theta S_2(t_2)Y^2 - [1 + \theta S_2(t_2)]\,Y + \{p + (1-p)F_1(t_1)\,[1 + \theta S_1(t_1)S_2(t_2)]\} = 0.$$

By definition, the discriminant of the quadratic equation is larger than zero and the smallest root lies within the interval $[0, 1]$.

In the three simulation settings we apply this with F_1 and F_2 the cumulative distribution functions for the Weibull distributions in (12). Details on (14) – (16) are provided in Appendix B of the Supplement.

In general, we generate data (z_i, δ_i, t_{2i}), given the aforementioned copula functions, and for different values of θ corresponding to Kendall's τ equal to 0.1, 0.2 and 0.5 (see Table 1). More specifically, $M = 500$ datasets with sample size $n = 250$ are generated under each scenario. Simulation results based on $M = 100$ datasets of sample size $n = 500$ are presented in Appendix C of the Supplement.

Table 1. Choices for θ in the simulation study and corresponding Kendall's τ values expressing the bivariate association in the data.

Copula function	Kendall's τ	Value	θ
Clayton	$\frac{\theta}{\theta+2}$	0.10	0.22
		0.20	0.50
		0.50	2.00
Gumbel	$\frac{\theta-1}{\theta}$	0.10	1.11
		0.20	1.25
		0.50	2.00
FGM[a]	$\frac{4\theta}{18}$	0.10	0.45
		0.20	0.90

[a]Given that $0 \leq \theta \leq 1$, we have $\tau \in [0, 2/9]$ (i.e., only weak dependence can be captured)

8.2. Simulation results

In Figure 1 we graphically depict the estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$, based on the algorithm presented in Appendix A of the Supplement, as a function of p for t_1 and t_2 equal to the median of the marginal Weibull distributions, i.e., $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$ for $i = 1, 2$. More specifically, simulation results are based on the Clayton copula used to generate 500 simulation sets with sample size equal to 250 observations and $\theta = 0.22$ (upper panels), $\theta = 0.50$ (middle panels) and $\theta = 2.00$ (lower panels). In the left panels we show the results under moderate censoring whereas results under heavy censoring are presented in the right panels. Similar to the estimation of the Kaplan-Meier estimator in a univariate setting, $\widetilde{Q}(p \mid t_1, t_2)$ can only be estimated for p-values for which

$$A^{-1/\theta}(p \mid t_1, t_2, \theta) \leq F_1[z_{(j)}],$$

with $z_{(j)}$ the largest uncensored observation (i.e., $\delta_{(j)} = 1$). Consequently, this upper bound is determined by the amount of censoring and similar inequalities can be obtained for other copula functions (see Appendix C of the Supplement for more details). Given the fact that the bound depends on the simulation run, we show the performance of the estimator for the minimum of these boundary values over the various simulation runs. The lower bound decreases with decreasing sample size and increasing censoring percentage. In each of the simulation runs a local cross-validation bandwidth selector as described in Geerdens et al. [17] is used to determine the step sizes of the Beran estimator.

In general, the averaged estimated quantiles $\widetilde{Q}_n(p \mid t_1, t_2)$ (green dashed lines) are close to the true quantiles (black solid line) for the entire range of possible p-values determined by the inequalities mentioned above. The simulation-based variability increases with increasing censoring percentage. Similar graphs for the Gumbel and FGM copula settings are presented in Appendix C of the Supplement.

9. Data application

We consider data on primary biliary cirrhosis (PBC) of the liver, a rare but fatal chronic liver condition of unknown cause that has been recognised since at least 1851 and that was renamed *primary biliary cholangitis* in 2014 [18] given the fact that liver cirrhosis is only a possible feature which arises when the disease is in an advanced stage. This autoimmune disease leads to the destruction of the small bile ducts in the liver and progression is slow, eventually leading to cirrhosis and liver decompensation.

Between January 1974 and May 1984, the Mayo Clinic conducted a double-blinded randomized clinical trial in PBC patients, comparing the survival time in patients treated with the drug D-penicillamine (DPCA) and of those in a placebo group. In total, 424 PBC patients met the eligibility criteria of the trial and 312 out of the 424 eligible patients agreed to participate in the trial. Next to the date of randomization, a large number of biochemical, serological and histological parameters were recorded for the clinical trial patients. For the additional 112 patients that did not participate in the trial, basic measurements were recorded and these patients agreed to follow-up for survival. Six of the later ones were lost to follow-up shortly after diagnosis, hence, were excluded from the data. The data analysis presented here included data on 418 study participants. The number of days between the start of the study and the earlier

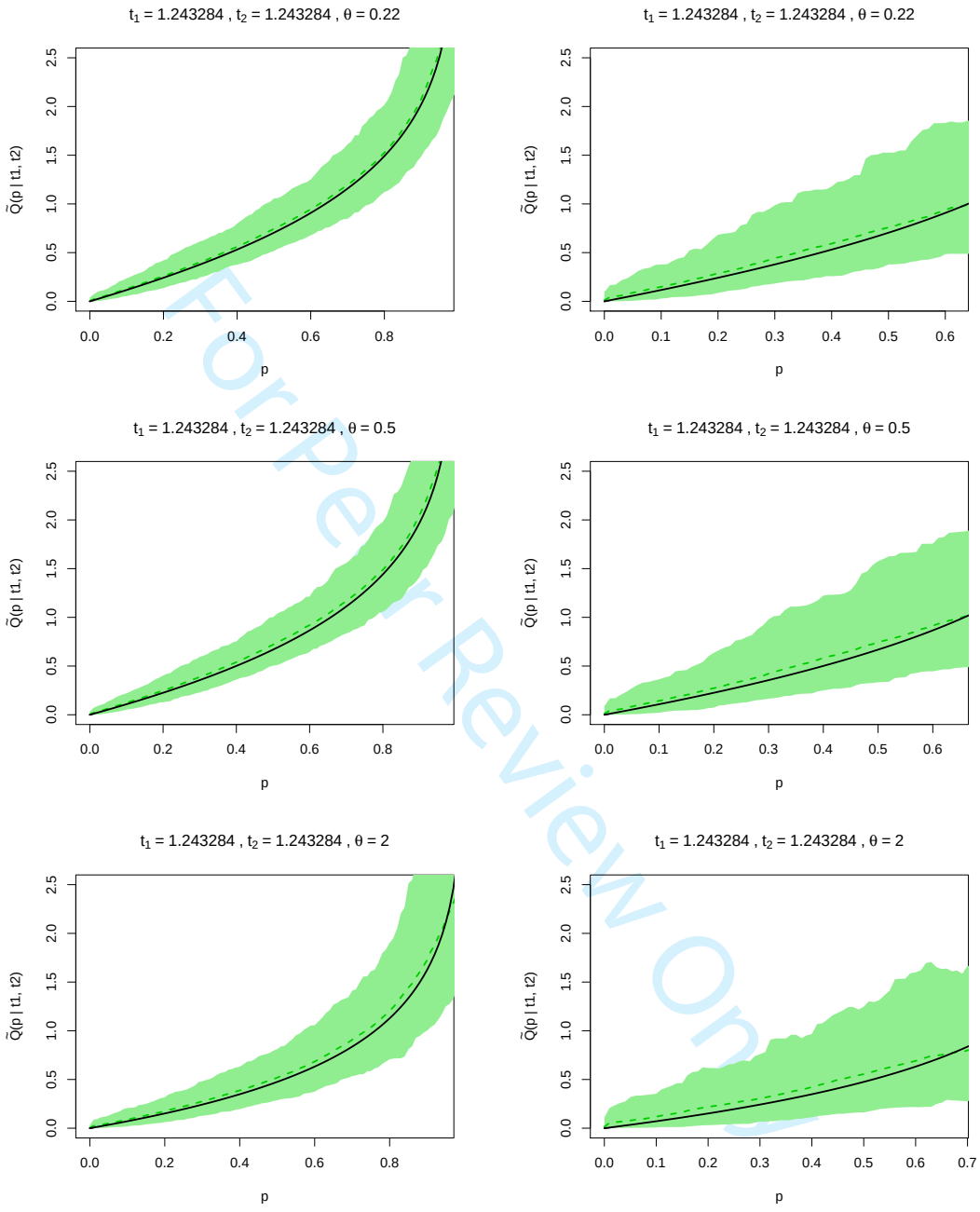


Figure 1. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 500 simulation sets of sample size 250 under the Clayton copula setting with $\theta = 0.22$ (upper panels), $\theta = 0.50$ (middle panels) and $\theta = 2.00$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

of death or transplantation is represented by T_1. Next to the survival time the serum bilirubin (expressed in mg/dl) for each of the patients is recorded.

In Figure 2, we show the survival times for PBC patients in the different treatment groups (DPCA, placebo, unrandomized patient) in relation to the serum bilirubin level, with the type of dot indicating whether the observation is right-censored or not (i.e., combining all-cause death or liver transplantation). As previously investigated by Fleming and Harrington [6], despite the immunosuppresive properties of DPCA, the survival time distributions of the DPCA group and the placebo group are similar. Also the survival time distribution of the unrandomized group is comparable to these of the DPCA and placebo groups. Hence, in this exercise, we will combine data from all patients to investigate the relationship between survival time T_1 and serum bilirubin levels prior to study entry. More specifically, we estimated the median residual lifetime for patients with a serum bilirubin level smaller than 3.4 (the 75% percentile of all serum bilirubin values) as compared to patients with levels exceeding this threshold. Needless to say, a stratified analysis could be conducted to study the median residual lifetimes in DPCA and placebo (+ unrandomized patients) separately (not shown here).

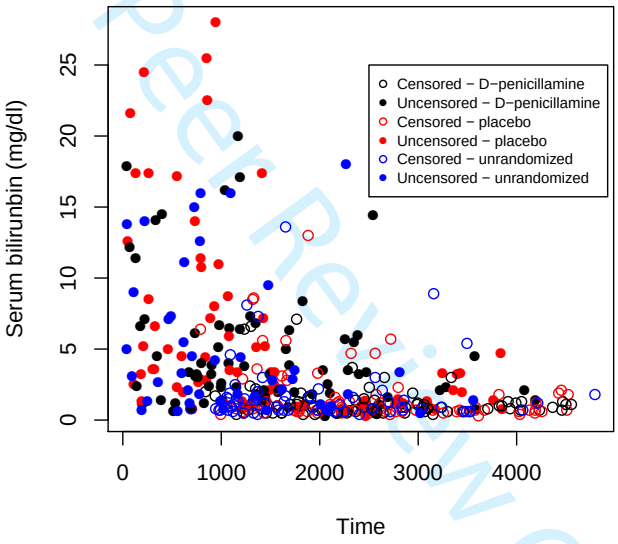


Figure 2. Observed survival time T_1 in relation with the serum bilirubin level (in mg/dl) for PBC patients in the DPCA (black dots), placebo (red dots) or unrandomized patients (blue dots). Open and closed dots represent censored and uncensored observations, respectively.

In Figure 3 the median residual lifetime is graphically depicted for patients with a serum bilirubin level less or equal than 3.4 mg/dl (left panel) and for patients with a level exceeding 3.4 mg/dl (right panel). More specifically, the black solid lines present the estimated median residual lifetimes based on the PBC data. Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based quantile functions are displayed using red solid lines. In general, the median residual lifetime is substantially lower for patients with a high serum bilirubin level ($T_2 > 3.4$) as compared to patients with a low ($T_2 \leq 3.4$) level. This confirms earlier findings regarding the prognostic performance of the serum bilirubin level [6].

The estimated ratio of the median residual lifetimes for these two patient groups is shown in Figure 4 (see Appendix D of the Supplement for more details). Given the

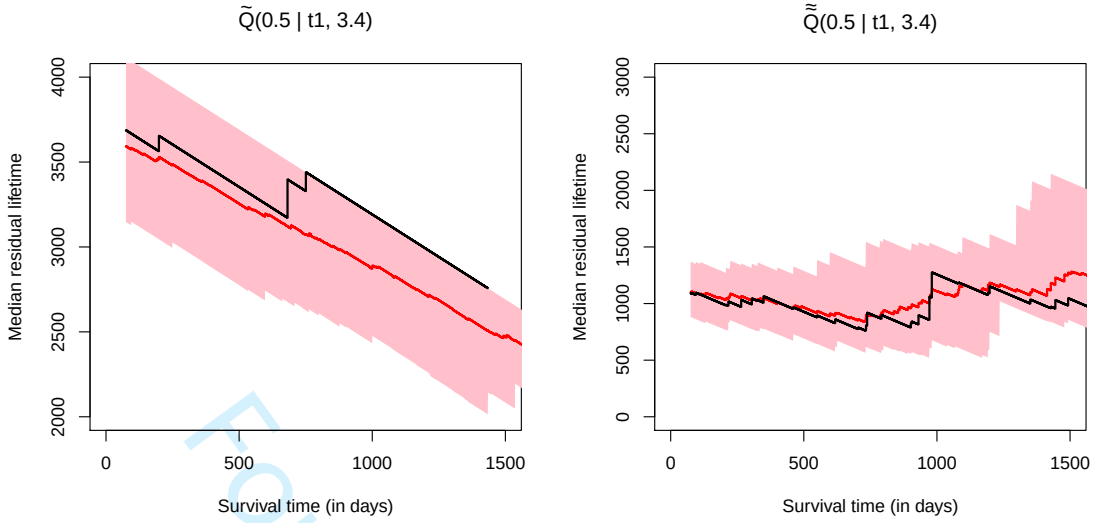


Figure 3. Estimated median residual lifetime (black solid lines) for patients with a serum bilirubin level less or equal than 3.4 mg/dl (left panel) and for patients with a serum bilirubin level exceeding 3.4 mg/dl (right panel). Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based median residual lifetime are represented by red solid lines.

fact that the ratio is well below one, we can conclude that patients with a high serum bilirubin levels have a median residual lifetime which is much lower than the median residual lifetime for patients with smaller bilirubin levels.

10. Discussion

In this manuscript, we present, for right-censored time-to-event data, novel nonparametric estimators for the quantiles of the conditional residual lifetime distribution $\widetilde{Q}(p \mid t_1, t_2)$ and $\widetilde{\widetilde{Q}}(p \mid t_1, t_2)$ based on plug-in estimators for the conditional cumulative distribution functions $\widetilde{F}_{t_2}(t_1) = P(T_1 \leq t_1 \mid T_2 \leq t_2)$, $\widetilde{\widetilde{F}}_{t_2}(t_1) = P(T_1 \leq t_1 \mid T_2 > t_2)$ and their corresponding quantiles.

A key ingredient in the estimators of the quantiles of the conditional residual lifetimes is the Beran estimator [11]. The fact that the Beran estimator relies on kernel-based Nadaraya-Watson weights implies the selection of optimal local or global bandwidths. In the implementation considered in the simulation study, we use the Epanechnikov kernel and local bandwidths are selected using cross-validation based on the approach proposed by Geerdens et al. [17].

The finite sample performance of our estimators is demonstrated in a simulation study. Although attention is mainly confined to the estimator for $\widetilde{Q}(p \mid t_1, t_2)$, similar results can be produced for the estimator $\widetilde{\widetilde{Q}}_n(p \mid t_1, t_2)$.

Note that, in the simulation procedure, we use the smallest observed $Z_{(l_2)}$-value for which

$$\widetilde{F}^{-1}_{t_2,n}[p + (1-p)\widetilde{F}_{t_2,n}(t_1)] \leq Z_{(l_2)}$$

to estimate the p-th quantile of the conditional residual lifetime for survival time

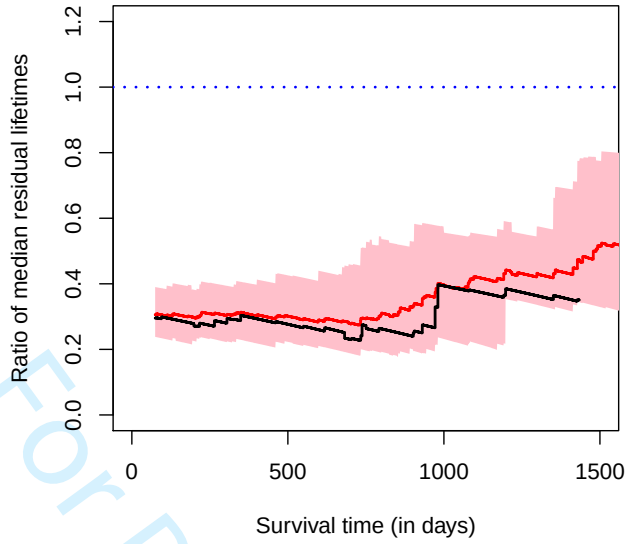


Figure 4. Ratio of estimated median residual lifetimes (black solid lines) for patients with a serum bilirubin level less or equal than 3.4 mg/dl and for patients with a serum bilirubin level exceeding 3.4 mg/dl. Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based ratio of median residual lifetimes is represented by a red solid line.

t_1, i.e., $-t_1 + Z_{(l_2)}$. This implies an overestimation of the quantile function, which asymptotically vanishes for $n \to \infty$. Alternatively, one could define the estimated quantile as a function of the largest ordered $Z_{(l_1)}$ for which $\delta_{(l_1)} = 1$ and

$$Z_{(l_1)} < \widetilde{F}^{-1}_{t_2,n}[p + (1-p)\widetilde{F}_{t_2,n}(t_1)],$$

i.e., $-t_1 + Z_{(l_1)}$. This provides an upper and lower bound estimator for the quantile function. Simulations provide very similar results with decreasing differences between these estimators asymptotically (not shown).

Our theoretical results provide an asymptotic variance expression $\widetilde{\sigma}^2_p(t_1, t_2)$ for $\widetilde{Q}(p \mid t_1, t_2)$ which, in theory, can be estimated based on plug-in estimators for $F_1(t_1)$, $F_2(t_2)$, $F(t_1, t_2)$, $\widetilde{F}_{t_2}(t_1)$ and $\widetilde{f}_{t_2}(t_1)$. Instead of estimating all the unknown quantities in the asymptotic variance, we use pointwise bootstrap-based confidence bounds in the data application. In the PBC data application, we illustrate the use of our novel estimators to estimate the median residual lifetime for patients with a high versus low bilirubin level. An additional data application based on lung cancer data is provided in the Supplementary Material.

Finally, the quantile functions of the conditional residual lifetime distributions, and their estimators, can be used to define a relative local association measure, similar to earlier association measures defined in terms of conditional hazards (see Appendix D of the Supplement). The advantage of such a local association measure over existing hazard-based methods is the direct interpretability in terms of the relative difference in, for example, median residual lifetime between two patients groups. Such interpretation is quintessential from a clinical perspective, for example, in the context of prognostic research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No specific funding was used by the author(s) to perform this work.

References

[1] Lin C, Zhang L, Zhou Y. Conditional quantile residual lifetime models for right censored data. Lifetime Data Anal. 2015;21:75–96.
[2] Liu Y, Lin C, Zhou Y. Nonparametric estimate of conditional quantile residual lifetime for right censored data. Statistics and Its Interface. 2019;12:61–70.
[3] Schmittlein DC, Morrison DG. The median residual life: a characterization theorem and an application. Operations Research. 1981;29:392–399.
[4] Joe H, Proshan F. Percentile residual life functions. Operations Research. 1984;32:668–678.
[5] Kayid M, Shafaei Moughabi M, Abouammoh AMM. A nonparametric estimator of bivariate quantile residual life model with application to tumor recurrence data set. Journal of Classification. 2020;37:237–253.
[6] Fleming TR, Harrington DP. Counting processes and survival analysis. John Wiley & Sons, Inc. 2005.
[7] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. 2000; Springer, New York.
[8] Abrams S, Janssen P, Swanepoel J et al. Nonparametric estimation of risk ratios for bivariate data. 2021. Submitted for publication.
[9] Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. The Annals of Statistics. 1994;22:1299–1327.
[10] Akritas MG, Van Keilegom I. Estimation of bivariate and marginal distributions with censored data. Journal of the Royal Statistical Society - Series B. 2003;65:457–471.
[11] Beran R. Nonparametric regression with randomly censored survival data. Technical report, Univ. of California, Berkeley. 1981.
[12] Van Keilegom I, Veraverbeke N. Estimation and bootstrap with censored data in fixed design nonparametric regression. Annals of the Institute of Statistical Mathematics. 1997;49:467–491.
[13] Lo S-H, Singh K. The product-limit estimator and the bootstrap: some asymptotic representations. Probability Theory and Related Fields. 1986;71:455–465.
[14] Gonzalez-Manteiga W, Cadarso-Suarez C. Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. Journal of Nonparametric Statistics. 1994;4:65–78.
[15] Ghosh JK. A new proof of the Bahadur representation of quantiles and an application. Annals of Mathematical Statistics. 1971;42:1957–1961.
[16] Gijbels I, Veraverbeke N. Weak asymptotic representations of quantiles of the product-limit estimator. Journal of Statistical Planning and Inference. 1988;18:151–160.
[17] Geerdens C, Acar EF, Janssen P. Conditional copula models for right-censored clustered event time data. Biostatistics. 2018;19:247–262.
[18] PBC Foundation (UK). PBC Name Change. https://www.pbcfoundation.org.uk/news/collettes-blog/pbc-name-change. Accessed November 1, 2021.

ORIGINAL ARTICLE

Supplementary Material for Quantiles of the conditional residual lifetime

S. Abrams[a,b] and P. Janssen[a,c] and N. Veraverbeke[a,c]

[a]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Data Science Institute (DSI), University of Hasselt, Diepenbeek, Belgium
[b]Global Health Institute (GHI), Family Medicine and Population Health (FAMPOP), University of Antwerp, Wilrijk, Belgium
[c]North-West University, Potchefstroom Campus, Potchefstroom, South Africa

ARTICLE HISTORY
Compiled November 1, 2021

CONTACT S. Abrams. Email: steven.abrams@uhasselt.be

Appendix A. Simulation algorithm

Given a specific simulation setting (including a copula function, parameter value θ and sample size n), we have information (Z_j, δ_j, T_{2j}), $j = 1, \ldots, n$ with $Z_j = \min(T_{1j}, C_j)$, δ_j the censoring indicator corresponding to event time T_{1j} and censoring time C_j, and T_{2j} the second random variable on which we will condition.

(i) Fix t_1 and t_2

(ii) Let $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$ denote the ordered Z_j, $j = 1, \ldots, n$ as defined previously; let $\delta_{(j)}$ denote the censoring indicator attached to $Z_{(j)}$

(iii) For a fixed i $(i = 1, \ldots, n)$ obtain the Beran estimator $F_n(t_1|t_{2i})$ for the conditional distribution function $F(t_1|t_{2i})$, i.e.,

$$F_n(t_1|t_{2i}) = \sum_{j=1}^{n} W_{nj}(t_{2i}) I(Z_{(j)} \leq t_1),$$

where, see, e.g., Van Keilegom and Veraverbeke [3],

$$W_{nj}(t_{2i}) = \delta_{(j)} \frac{w_{n(j)}(t_{2i}, h_n)}{1 - \sum_{k=1}^{j-1} w_{n(k)}(t_{2i}, h_n)} \prod_{l=1}^{j-1} \left(1 - \frac{w_{n(l)}(t_{2i}, h_n)}{1 - \sum_{k=1}^{l-1} w_{n(k)}(t_{2i}, h_n)} \right)^{\delta_{(l)}} \qquad (j < n),$$

$$W_{nn}(t_{2i}) = \prod_{l=1}^{n-1} \left(1 - \frac{w_{n(l)}(t_{2i}, h_n)}{1 - \sum_{k=1}^{l-1} w_{n(k)}(t_{2i}, h_n)} \right)^{\delta_{(l)}},$$

and where $w_{n(j)}(t_{2i}, h_n)$ are the previously defined weights $w_{nj}(t_{2i}, h_n)$ corresponding to the ordering of the $Z_{(j)}$ as defined in (ii) above. Note that the Beran estimator only jumps at $Z_{(j)}$ values with $\delta_{(j)} = 1$ and that in case of no right censoring, $W_{nj}(t_{2i}) = w_{n(j)(t_{2i}, h_n)}$ such that $F_n(t_1|t_{2i})$ simplifies to the kernel estimator of Stone [4] for the conditional distribution function $F(t_1 \mid t_2)$.

(iv) As a result, we obtain an $(n \times n)$-matrix $\boldsymbol{W} = \{W_{nj}(t_{2i})\}_{j=1,\ldots,n;i=1,\ldots,n}$.

(v) Define, for $j = 1, \ldots, n$,

$$\widetilde{W}_{nj}(t_2) = \frac{\sum_{i=1}^{n} W_{nj}(T_{2i}) I(T_{2i} \leq t_2)}{\sum_{i=1}^{n} I(T_{2i} \leq t_2)}.$$

We then have

$$\widetilde{F}_{t_2,n}(t_1) = \sum_{j=1}^{n} \widetilde{W}_{nj}(t_2) I(Z_{(j)} \leq t_1).$$

Similar quantities can be obtained for $\widetilde{\widetilde{W}}_{nj}(t_2)$ and $\widetilde{\widetilde{F}}_{t_2,n}(t_1)$. See Appendix B.2. for more details on these formulas and the derivation of the expression for $\widetilde{F}_{t_2,n}(t_1)$ and $\widetilde{\widetilde{F}}_{t_2,n}(t_1)$.

(vi) Define, for $j = 1, \ldots, n$,

$$S_{nj}(t_2) = \sum_{l=1}^{j} \widetilde{W}_{nl}(t_2).$$

Obtain the largest l_1 and the smallest $l_2 > l_1 \in \{1, \ldots, n\}$, such that

$$S_{nl_1}(t_2) < p + (1-p)\widetilde{F}_{t_2,n}(t_1) \leq S_{nl_2}(t_2).$$

Then

$$Z_{(l_1)} < \widetilde{F}^{-1}_{t_2,n}[p + (1-p)\widetilde{F}_{t_2,n}(t_1)] \leq Z_{(l_2)},$$

where $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ are the ordered Z_j's (see (ii)). In the simulations, we define

$$\widetilde{F}^{-1}_{t_2,n}[p + (1-p)\widetilde{F}_{t_2,n}(t_1)] := Z_{(l_2)},$$

(vii) Finally, obtain

$$y = -t_1 + Z_{(l_2)}.$$

Appendix B. Additional derivations

B.1. Formulas for $\widetilde{Q}(p \mid t_1, t_2, \theta)$

Let $F(t_1, t_2) = C\left[F_1(t_1), F_2(t_2)\right]$, with $C \in \mathcal{C}_\theta$ (i.e., a one-parameter copula family). To obtain $\widetilde{Q}(p \mid t_1, t_2)$ we need to solve the following equation with respect to y

$$C(u(y), v) = pv + (1-p)C(u, v), \tag{B1}$$

where $u(y) = F_1(t_1+y)$, $u = u(0) = F_1(t_1)$, $v = F_2(t_2)$. We use, as shorthand notation,

$$B(p \mid t_1, t_2, \theta) = pv + (1-p)C(u, v).$$

B.1.1. Clayton copula

For the Clayton copula, we have

$$C_C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}.$$

Consequently,

$$\left[C_C(u, v)\right]^{-\theta} = u^{-\theta} + v^{-\theta} - 1.$$

Equation (B1) renders

$$\begin{aligned}
&u(y)^{-\theta} + v^{-\theta} - 1 = \left[B(p \mid t_1, t_2, \theta)\right]^{-\theta} \\
&\Leftrightarrow u(y)^{-\theta} = \left[B(p \mid t_1, t_2, \theta)\right]^{-\theta} - v^{-\theta} + 1 \\
&\Leftrightarrow F_1(t_1 + y) = \left\{\left[B(p \mid t_1, t_2, \theta)\right]^{-\theta} - v^{-\theta} + 1\right\}^{-1/\theta} \\
&\Leftrightarrow y = -t_1 + F_1^{-1}\left(\left\{\left[B(p \mid t_1, t_2, \theta)\right]^{-\theta} - v^{-\theta} + 1\right\}^{-1/\theta}\right) \\
&\Leftrightarrow y \equiv -t_1 + F_1^{-1}\left[A^{-1/\theta}(p \mid t_1, t_2, \theta)\right].
\end{aligned}$$

B.1.2. Gumbel copula

For the Gumbel copula, we have

$$C_G(u, v) = \exp\left(-\left\{\left[-\log(u)\right]^\theta + \left[-\log(v)\right]^\theta\right\}^{1/\theta}\right).$$

Hence,

$$\left\{-\log\left[C_G(u, v)\right]\right\}^\theta = \left[-\log(u)\right]^\theta + \left[-\log(v)\right]^\theta.$$

Therefore, equation (B1) can be rewritten as

$$\begin{aligned}
&\{-\log[u(y)]\}^\theta + [-\log(v)]^\theta = \{-\log\left[B(p \mid t_1, t_2, \theta)\right]\}^\theta \\
&\Leftrightarrow \log[u(y)] = -\left(\{-\log\left[B(p \mid t_1, t_2, \theta)\right]\}^\theta - [-\log(v)]^\theta\right)^{1/\theta} \\
&\Leftrightarrow F_1(t_1 + y) = \exp\left[-\left(\{-\log\left[B(p \mid t_1, t_2, \theta)\right]\}^\theta - [-\log(v)]^\theta\right)^{1/\theta}\right] \\
&\Leftrightarrow y = -t_1 + F_1^{-1}\left\{\exp\left[-\left(\{-\log\left[B(p \mid t_1, t_2, \theta)\right]\}^\theta - [-\log(v)]^\theta\right)^{1/\theta}\right]\right\}.
\end{aligned}$$

B.1.3. Farlie-Gumbel-Morgenstern copula

In contrast to Archimedean copulas, an analytical solution for equation (B1) is not guaranteed for non-Archimedean copulas like the Farlie-Gumbel-Morgenstern copula.

B.2. Formulas for $\widetilde{F}_{t_2,n}(t_1)$ and $\widetilde{\widetilde{F}}_{t_2,n}(t_1)$

$$\begin{aligned}
\widetilde{F}_{t_2,n}(t_1) &= \frac{F_n(t_1, t_2)}{F_{2n}(t_2)} = \frac{\sum_{i=1}^n F_n(t_1|T_{2i}) I(T_{2i} \le t_2)}{\sum_{i=1}^n I(T_{2i} \le t_2)} \\
&= \frac{1}{\sum_{i=1}^n I(T_{2i} \le t_2)} \sum_{i=1}^n \left\{\sum_{j=1}^n W_{nj}(T_{2i}) I(Z_{(j)} \le t_1)\right\} I(T_{2i} \le t_2) \\
&= \sum_{j=1}^n \left\{\frac{\sum_{i=1}^n W_{nj}(T_{2i}) I(T_{2i} \le t_2)}{\sum_{i=1}^n I(T_{2i} \le t_2)}\right\} I(Z_{(j)} \le t_1) \\
&=: \sum_{j=1}^n \widetilde{W}_{nj}(t_2) I(Z_{(j)} \le t_1).
\end{aligned}$$

$$\begin{aligned}
\widetilde{\widetilde{F}}_{t_2,n}(t_1) &= \frac{\sum_{i=1}^n F_n(t_1|T_{2i}) I(T_{2i} > t_2)}{\sum_{i=1}^n I(T_{2i} > t_2)} \\
&= \frac{1}{\sum_{i=1}^n I(T_{2i} > t_2)} \sum_{i=1}^n \left\{\sum_{j=1}^n W_{nj}(T_{2i}) I(Z_{(j)} \le t_1)\right\} I(T_{2i} > t_2) \\
&= \sum_{j=1}^n \left\{\frac{\sum_{i=1}^n W_{nj}(T_{2i}) I(T_{2i} > t_2)}{\sum_{i=1}^n I(T_{2i} > t_2)}\right\} I(Z_{(j)} \le t_1) \\
&=: \sum_{j=1}^n \widetilde{\widetilde{W}}_{nj}(t_2) I(Z_{(j)} \le t_1).
\end{aligned}$$

Appendix C. Additional simulation results

C.1. Clayton copula

In Figure C1 we show the estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p for t_1 and t_2 equal to the median of the marginal Weibull distributions, i.e., $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$ for $i = 1, 2$. Simulation results are based on the Clayton copula used to generate 100 simulation sets with sample size equal to 500 observations and $\theta = 0.22$ (upper panels), $\theta = 0.50$ (middle panels) and $\theta = 2.00$ (lower panels). In the left panels we show the results under moderate censoring whereas results under heavy censoring are presented in the right panels. Clearly, the increase in sample size reduces the simulation-based uncertainty.

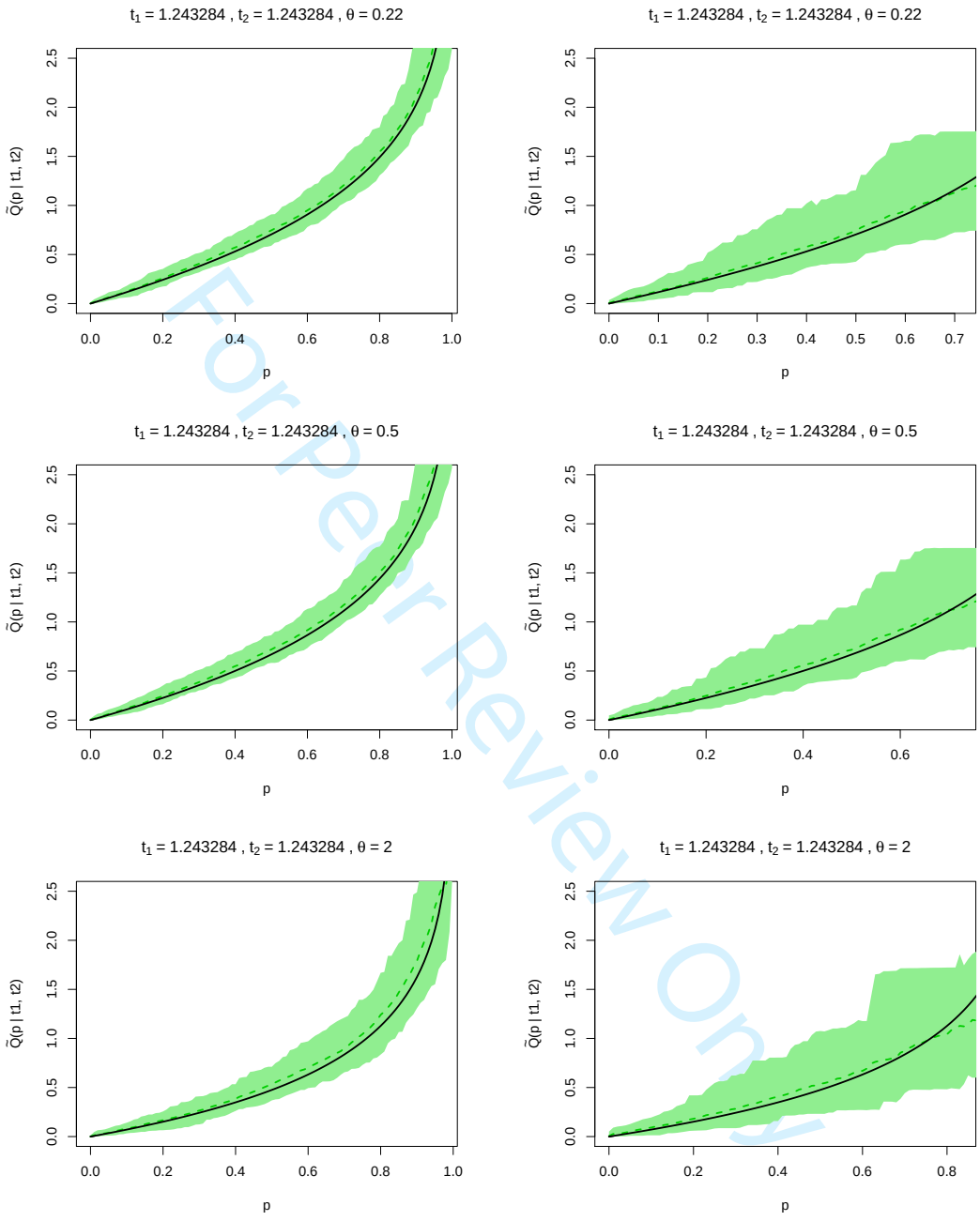


Figure C1. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 100 simulation sets of sample size 500 under the Clayton copula setting with $\theta = 0.22$ (upper panels), $\theta = 0.50$ (middle panels) and $\theta = 2.00$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

C.2. Gumbel copula

In Figure C2 we show the estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p for t_1 and t_2 equal to the median of the marginal Weibull distributions. Simulation results are based on the Gumbel copula used to generate 500 simulation sets with sample size equal to 250 observations and $\theta = 1.11$ (upper panels), $\theta = 1.25$ (middle panels) and $\theta = 2.00$ (lower panels). In the left panels we show the results under moderate censoring whereas results under heavy censoring are presented in the right panels. The same simulation scenarios are explored for an increased sample size of 500 observations per simulation run (see Figure C3).

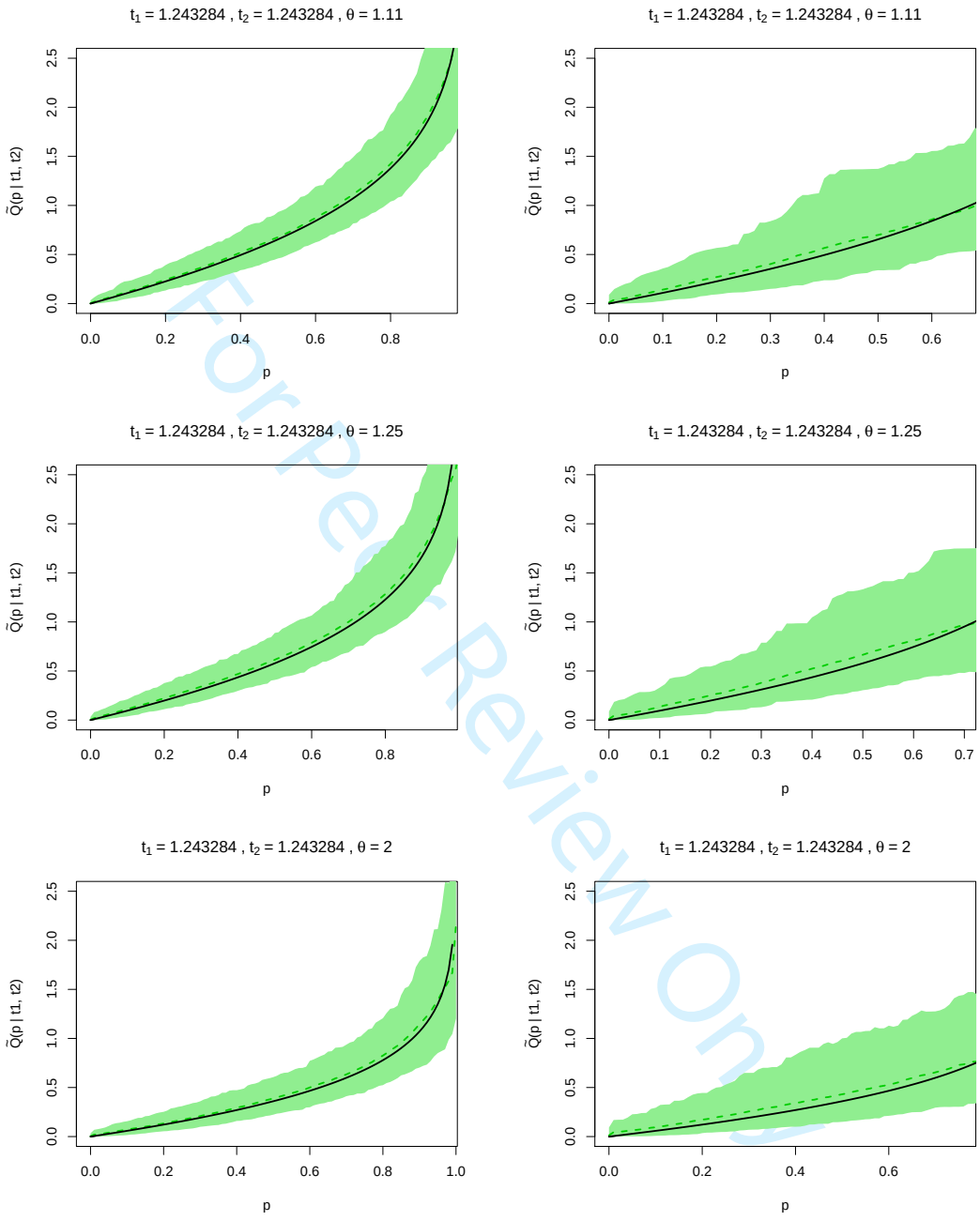


Figure C2. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 500 simulation sets of sample size 250 under the Gumbel copula setting with $\theta = 1.11$ (upper panels), $\theta = 1.25$ (middle panels) and $\theta = 2.00$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

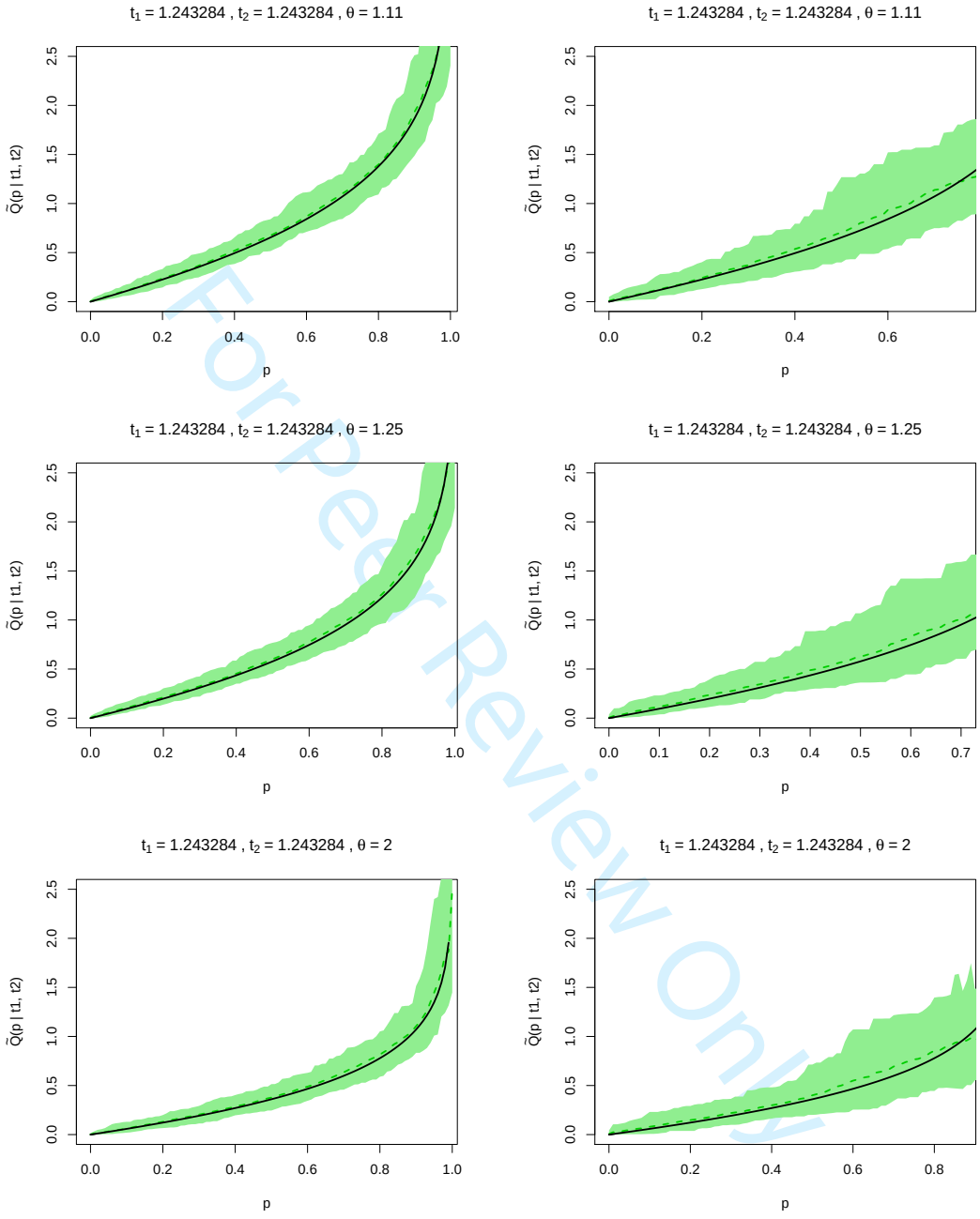


Figure C3. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 100 simulation sets of sample size 500 under the Gumbel copula setting with $\theta = 1.11$ (upper panels), $\theta = 1.25$ (middle panels) and $\theta = 2.00$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

C.3. Farlie-Gumbel-Morgenstern copula

In Figure C4 we show the estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p for t_1 and t_2 equal to the median of the marginal Weibull distributions. Simulation results are based on the FGM copula used to generate 500 simulation sets with sample size equal to 250 observations and $\theta = 0.45$ (upper panels) and $\theta = 0.90$ (lower panels). In the left panels we show the results under moderate censoring whereas results under heavy censoring are presented in the right panels. The same simulation scenarios are explored for an increased sample size of 500 observations per simulation run (see Figure C5).

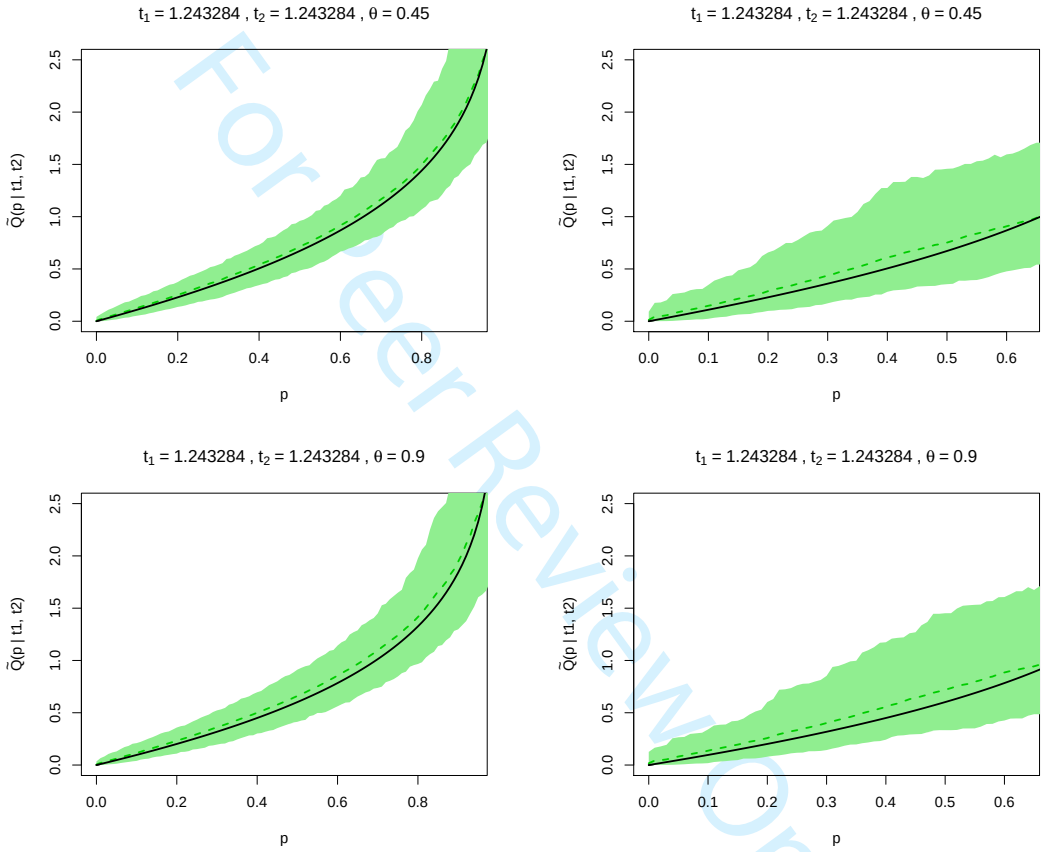


Figure C4. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 500 simulation sets of sample size 250 under the Farlie-Gumbel-Morgenstern copula setting with $\theta = 0.45$ (upper panels) and $\theta = 0.90$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

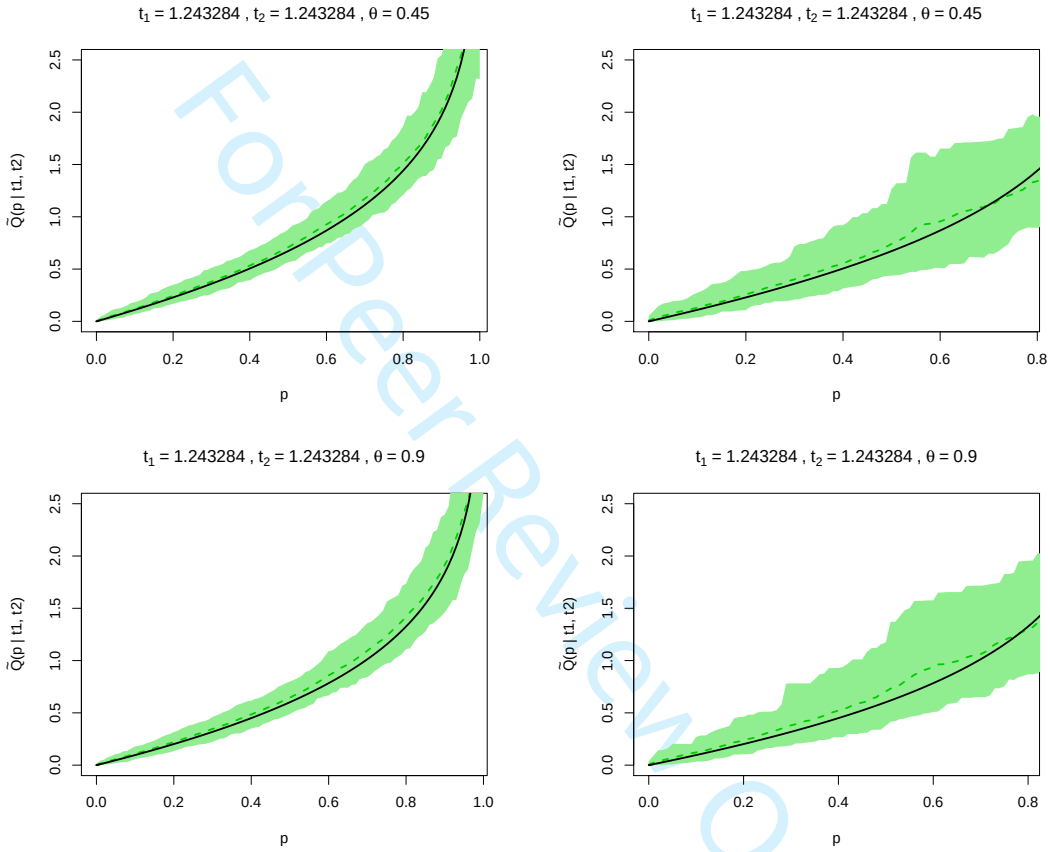


Figure C5. True quantile function $\widetilde{Q}(p \mid t_1, t_2)$ (black solid line) and averaged estimated quantile function $\widetilde{Q}_n(p \mid t_1, t_2)$ as a function of p (green dashed lines) for $t_i = d_i^{-1/s_i}[\ln(2)]^{1/s_i} \approx 1.243$, $i = 1, 2$ and based on 100 simulation sets of sample size 500 under the Farlie-Gumbel-Morgenstern copula setting with $\theta = 0.45$ (upper panels) and $\theta = 0.90$ (lower panels). Left panels show the results under moderate censoring and right panels under heavy censoring. Pointwise 95% confidence limits are shown as green shaded area.

Appendix D. Advanced lung cancer data application

As a second illustrative example, we consider the North Central Cancer Treatment Group (NCCTG) dataset, available in the R package *survival*, describing the survival of patients with advanced lung cancer [2]. More specifically, information is available for 228 patients suffering from lung cancer with T_1 the survival time between onset of lung cancer and death ($n_d = 165$) or censoring ($n_c = 63$). Next to performance scores representing how well the patient can perform usual daily activities, weight loss in the last six months before entering the study, denoted by T_2, was available.

In Figure D1 the median residual lifetime is graphically depicted for patients with a Karnofsky performance score, measured at a 0-100 scale and rated by the physician, less or equal than the median score of 80 points (left panel) and for patients with a score exceeding this median value (right panel). More specifically, the black solid lines present the estimated median residual lifetimes based on the data. Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based quantile functions are displayed using red solid lines. In general, the median residual lifetime increases substantially for patients with a high score ($T_2 > 80$) as compared to patients with a low ($T_2 \leq 80$) performance score. Hence, the figure provides graphical evidence that the Karnofsky score might be a useful predictor for the (residual) lifetime distribution. This confirms earlier findings regarding the prognostic performance of the score [2].

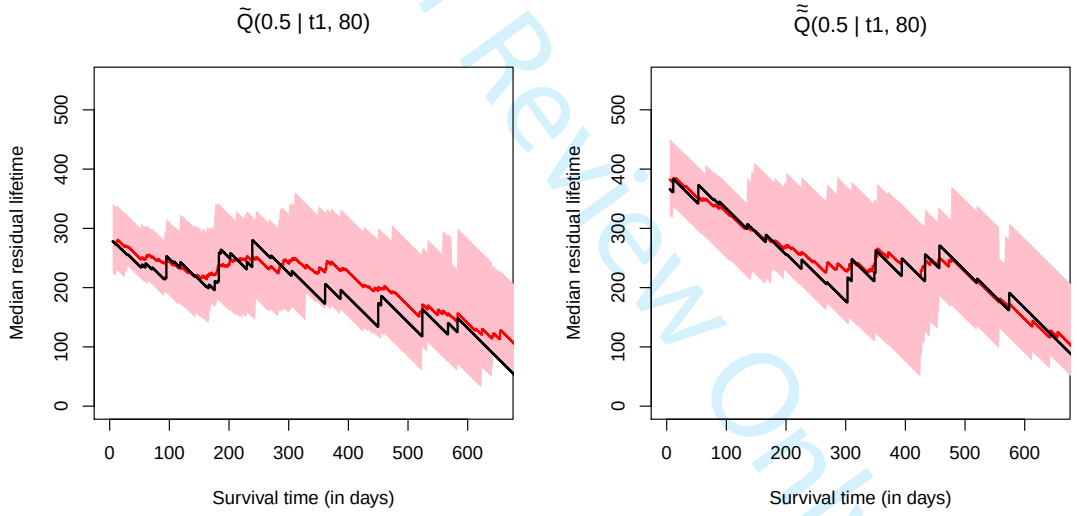


Figure D1. Estimated median residual lifetime (black solid lines) for patients with a Karnofsky performance score less or equal than the median score of 80 points (left panel) and for patients with a score exceeding the median score of 80 points (right panel). Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based median residual lifetime are represented by red solid lines.

In Abrams et al. [1], a risk measure was studied based on the ratio of two conditional hazard rate functions. An alternative for this association measure can now be proposed as follows:

$$R(t_1, t_2) = \frac{\tilde{\tilde{Q}}(p \mid t_1, t_2)}{\tilde{Q}(p \mid t_1, t_2)}.$$

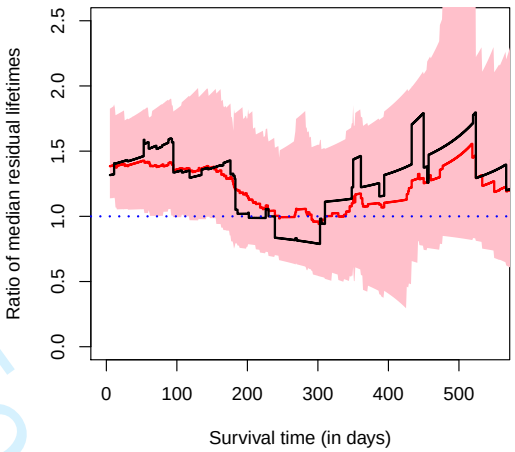


Figure D2. Ratio of estimated median residual lifetimes (black solid lines) for patients with a Karnofsky performance score exceeding the median score of 80 points and for patients with a score smaller or equal to this median value. Pointwise bootstrap-based 95% confidence bounds are shown as a pink shaded area and the average bootstrap-based ratio of median residual lifetimes is represented by a red solid line.

This ratio compares the p-th quantile of the residual lifetime of T_1 after t_1, in the group with $T_2 > t_2$ and the group with $T_2 \leq t_2$. If T_1 and T_2 are independent then

$$\widetilde{F}_{t_2}(t_1) = \widetilde{\widetilde{F}}_{t_2}(t_1) = F_1(t_2).$$

Consequently, $R(t_1, t_2) = 1$ under independence. A natural estimator for $R(t_1, t_2)$ is then

$$R_n(t_1, t_2) = \frac{\widetilde{\widetilde{Q}}_n(p \mid t_1, t_2)}{\widetilde{Q}_n(p \mid t_1, t_2)}.$$

Combining the asymptotic representations of Theorems 6 and 7, we can easily derive that

$$n^{1/2} \left[R_n(t_1, t_2) - R(t_1, t_2) \right]$$

is asymptotically normal. The ratio of the estimated median residual lifetimes for the abovementioned patient groups is shown in Figure D2. The estimated ratio of the median residual lifetimes for these patient groups is shown in Figure D2. The figure provides clear evidence that, for conditioning on the smaller survival times, the estimated ratio of the median residual lifetimes is larger than one, hence, patients with high scores have a higher median residual lifetime, at least for small survival times t_1.

A comparison (not shown) of median residual lifetimes based on age (younger versus older) or weight loss (low versus high) suggests that the predictive potential of age and weight loss is small.

References

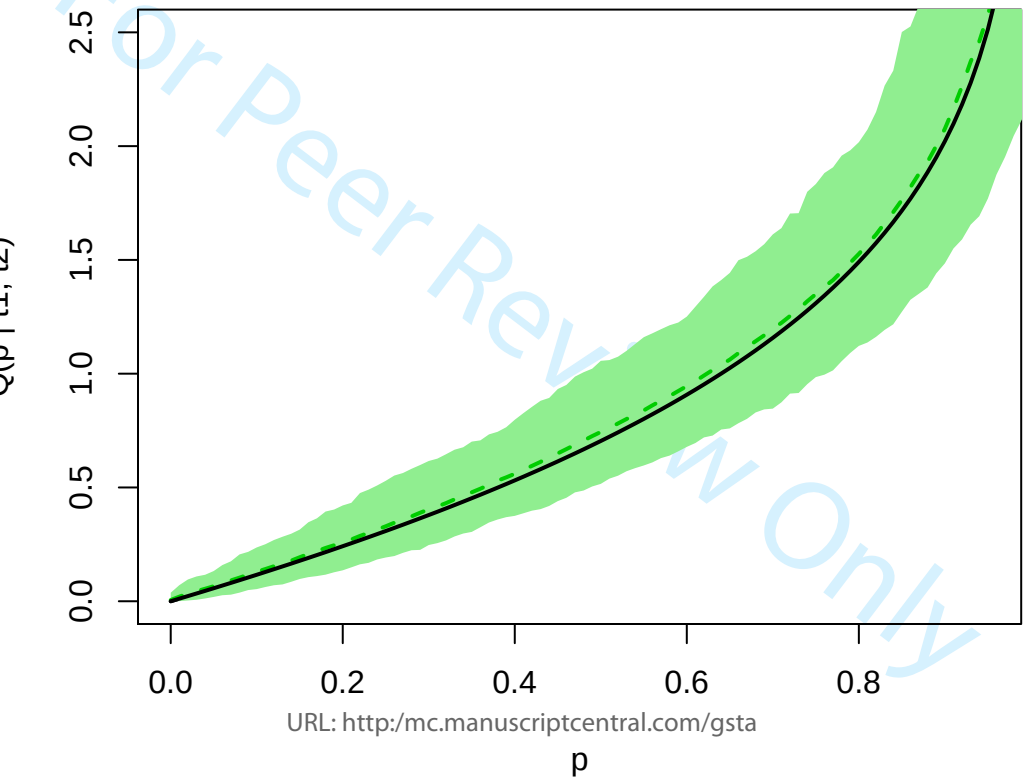
[1] Abrams S, Janssen P, Swanepoel J et al. Nonparametric estimation of risk ratios for bivariate data. 2021. Submitted for publication.

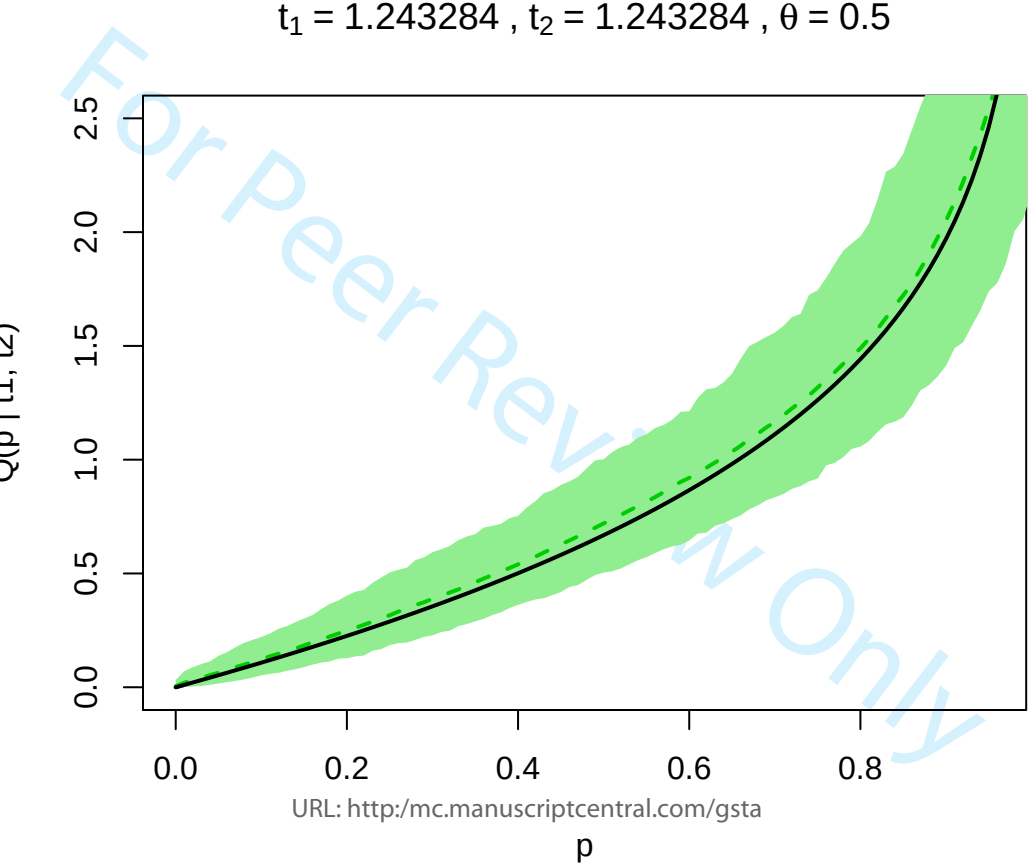
[2] Loprinzi CL, Laurie JA, Wieand HS et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 1994;12:601–607.

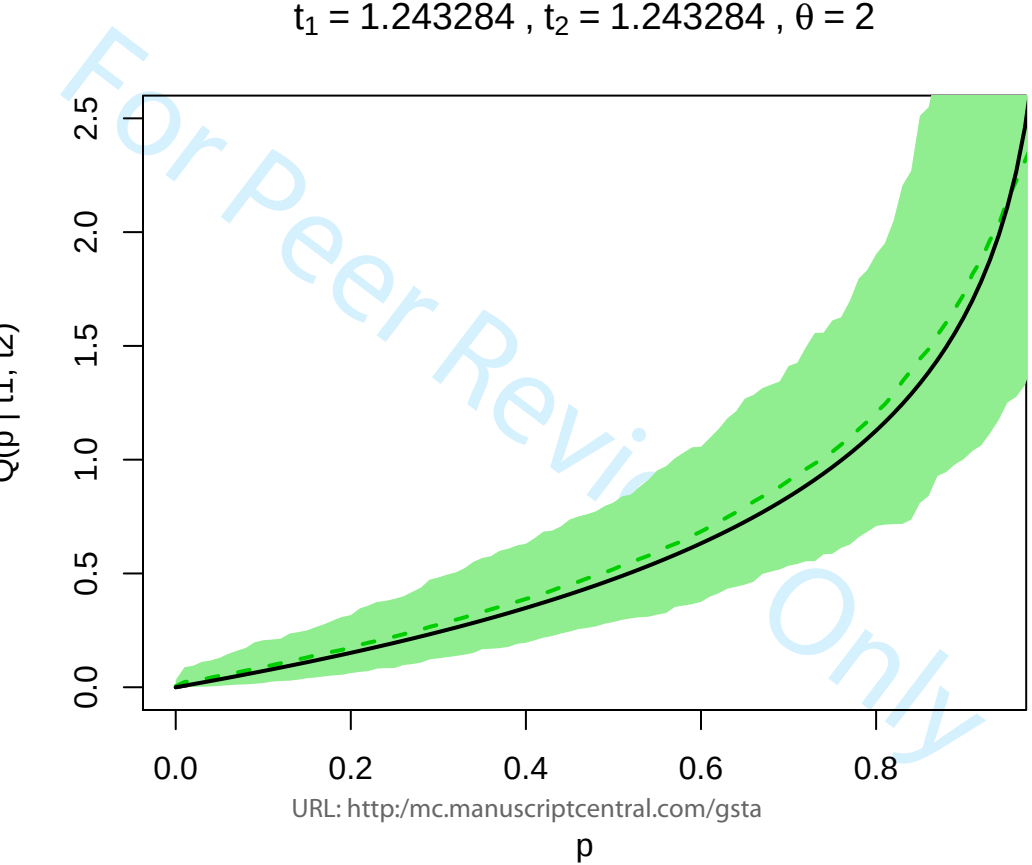
[3] Van Keilegom I. Nonparametric estimation of the conditional distribution in regression with censored data. Doctoral thesis. 1998. Available at https://ibiostat.be/publications/phd/ingridvankeilegom.pdf.

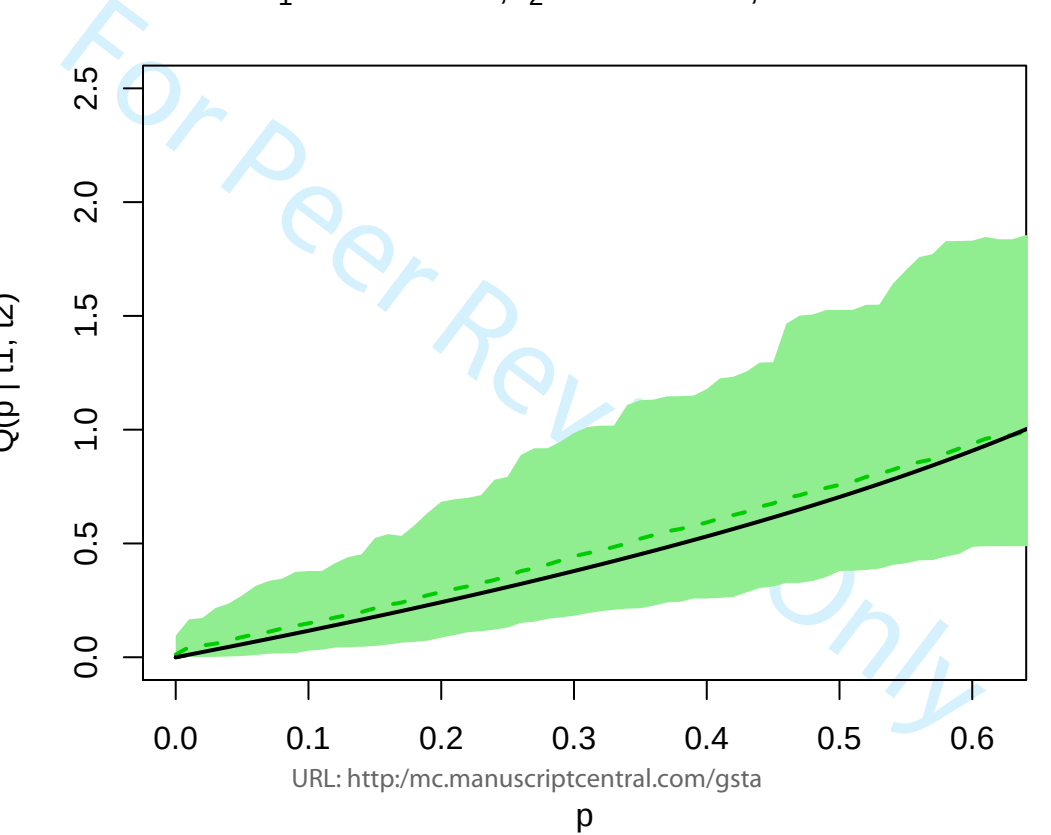
[4] Stone CJ. Consistent nonparametric regression. The Annals of Statistics. 1977;5:595–645.

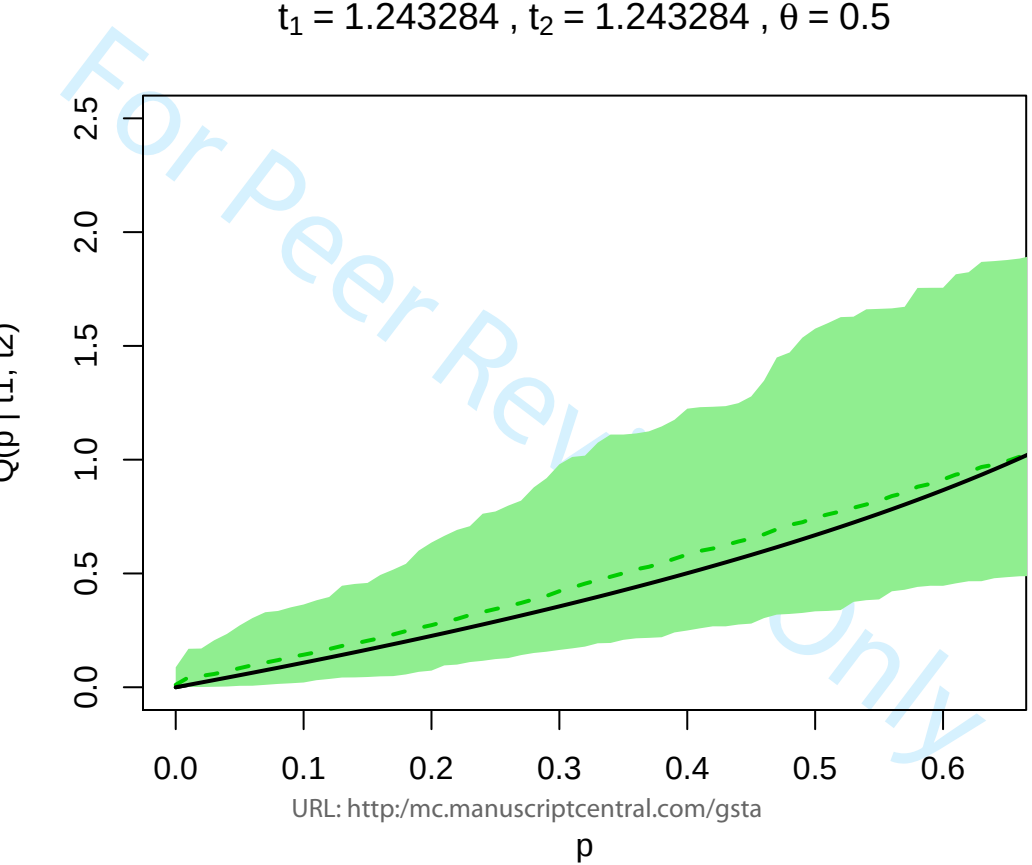
$t_1 = 1.243284$, $t_2 = 1.243284$, $\theta = 0.22$

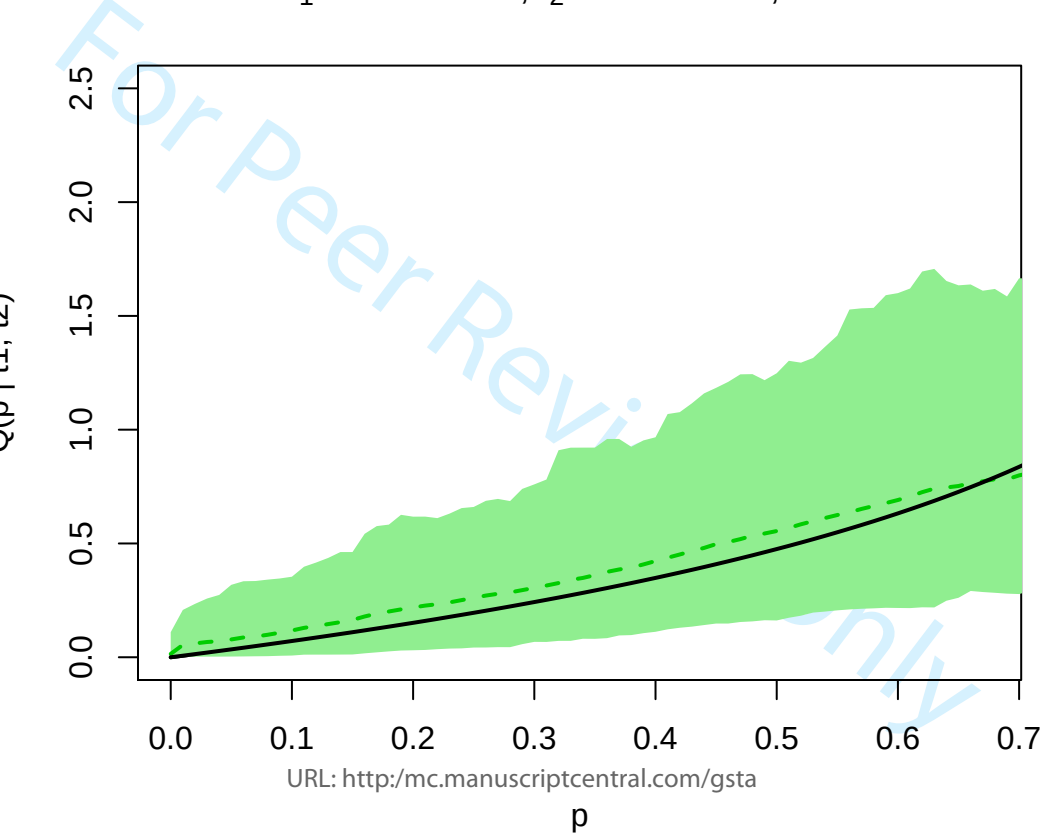


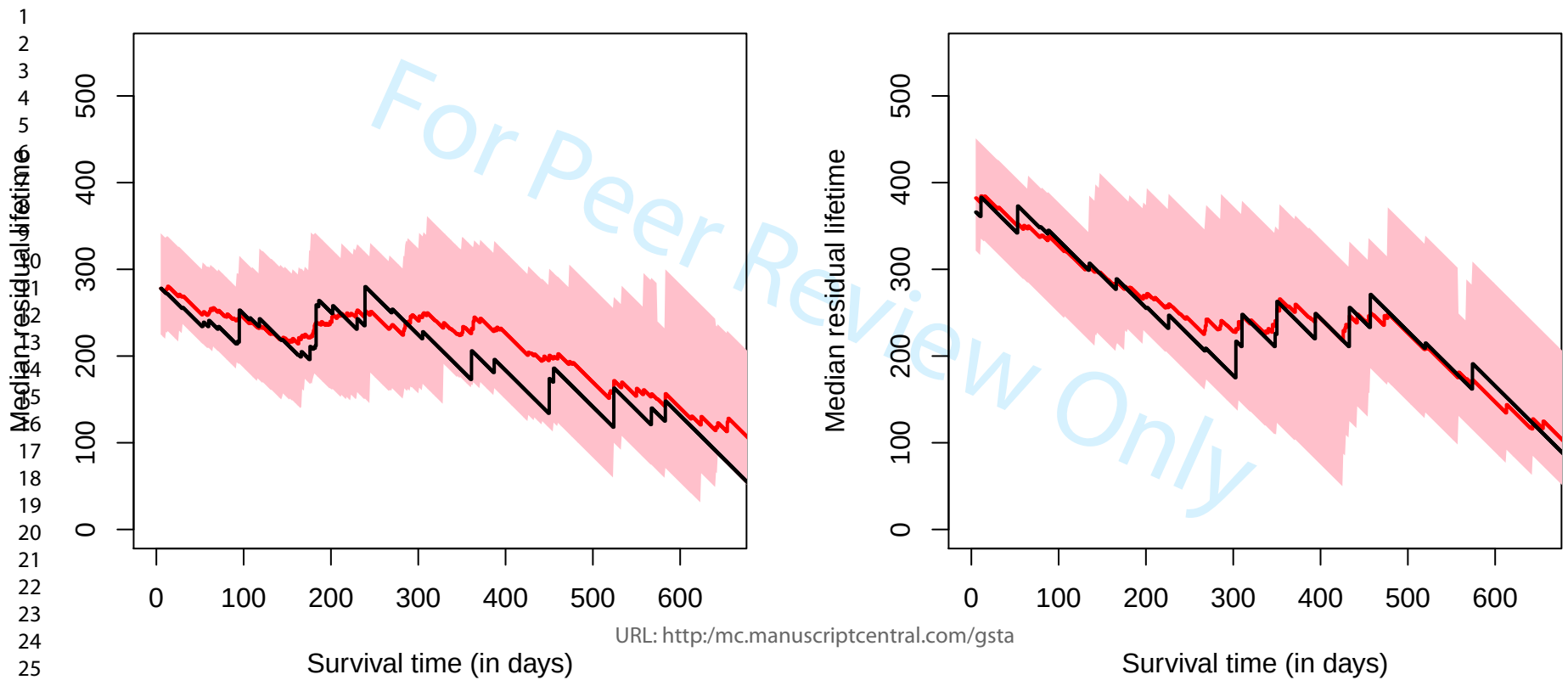


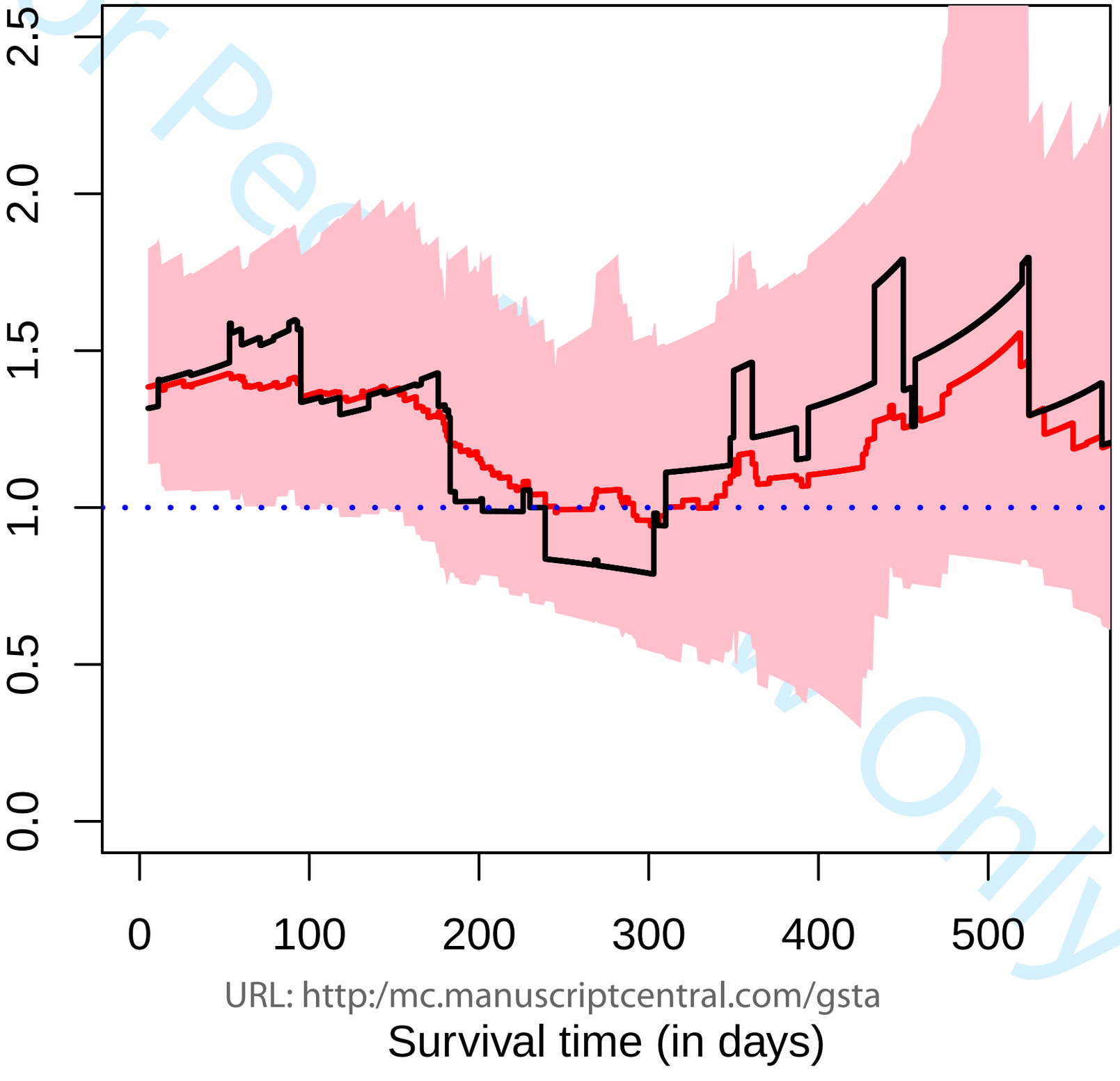


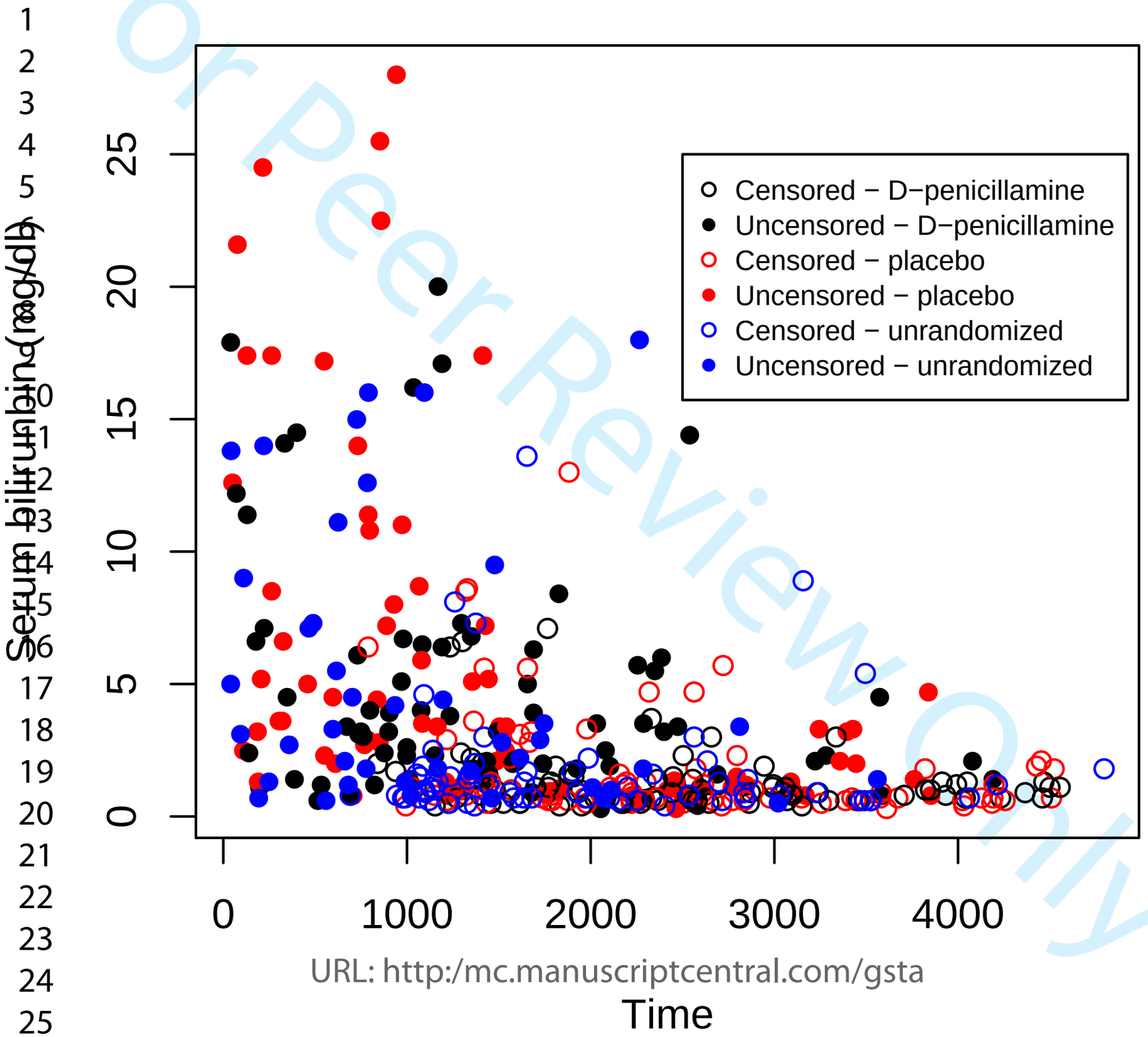


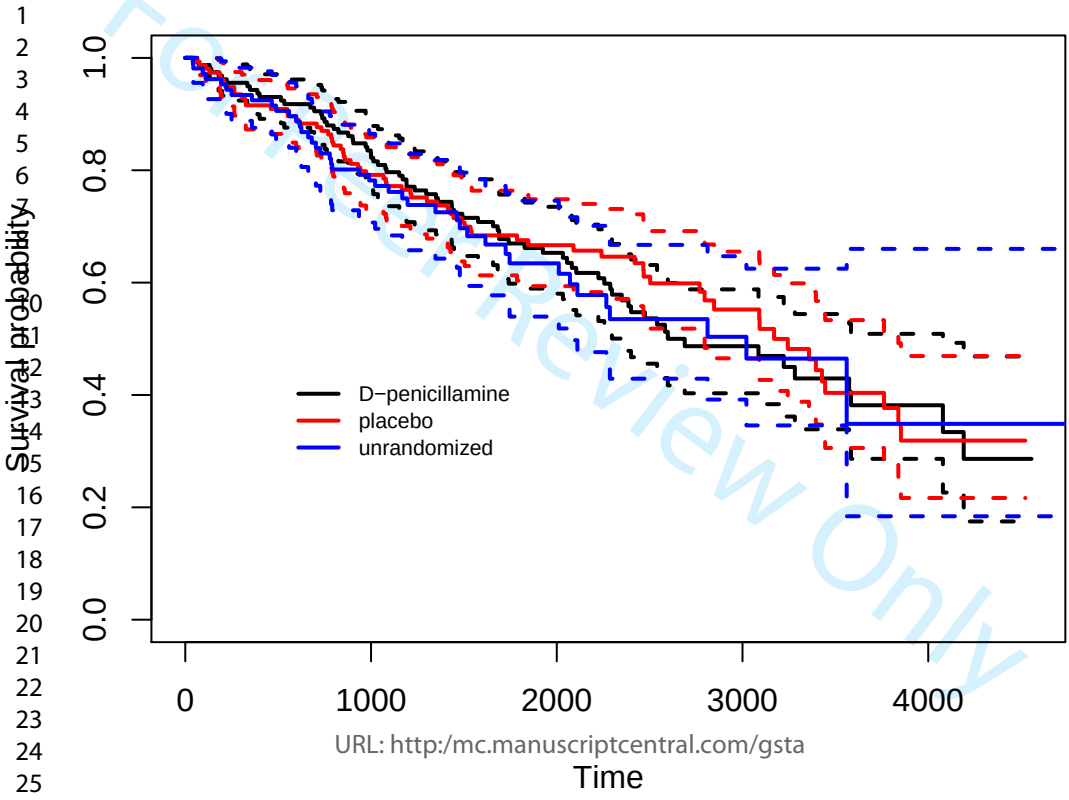


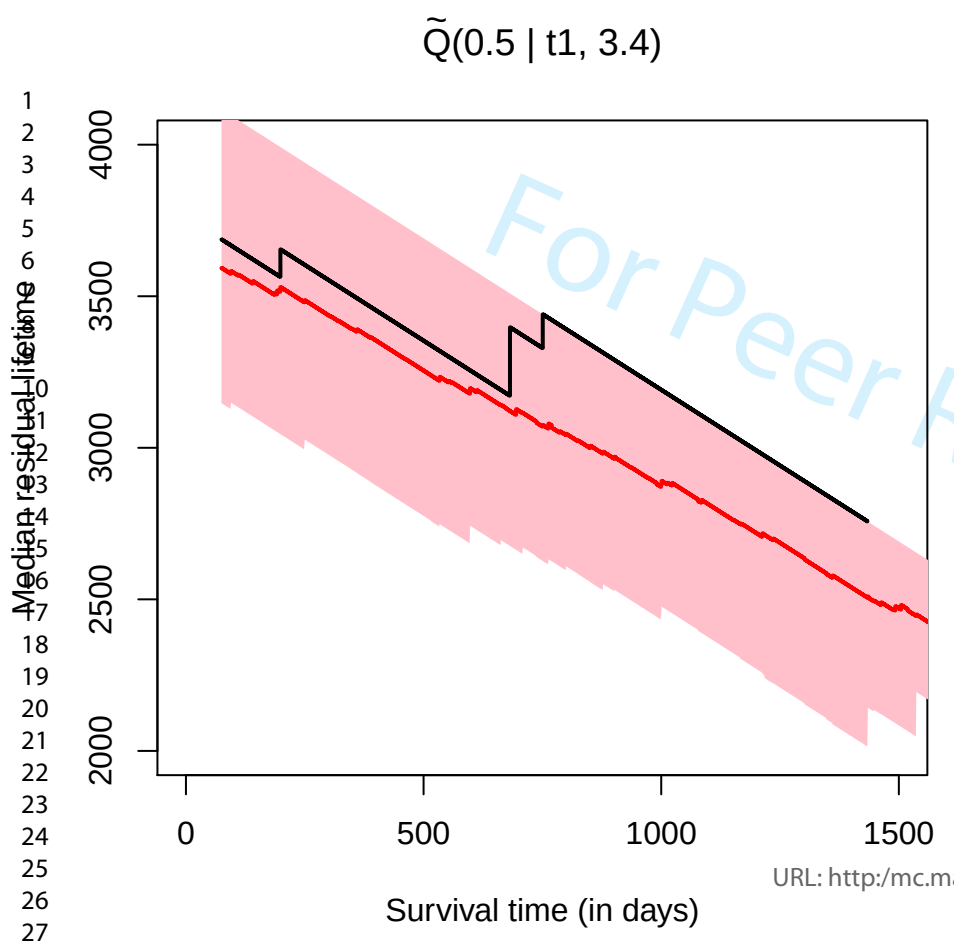


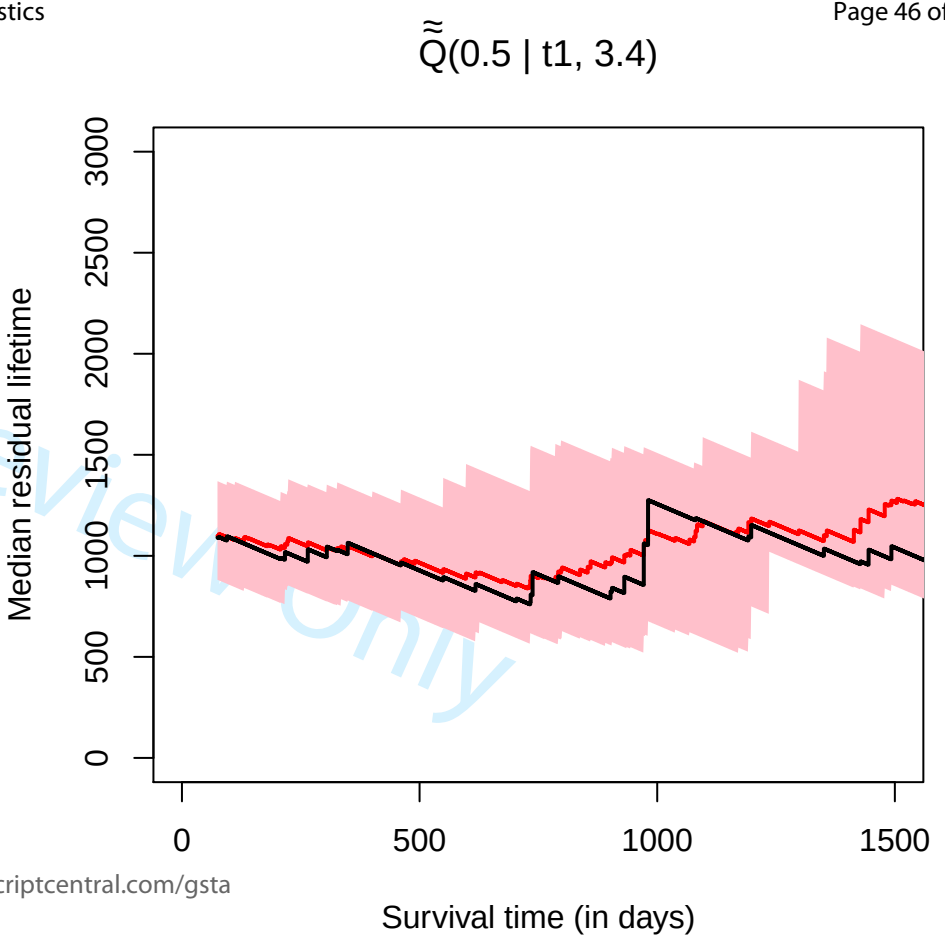


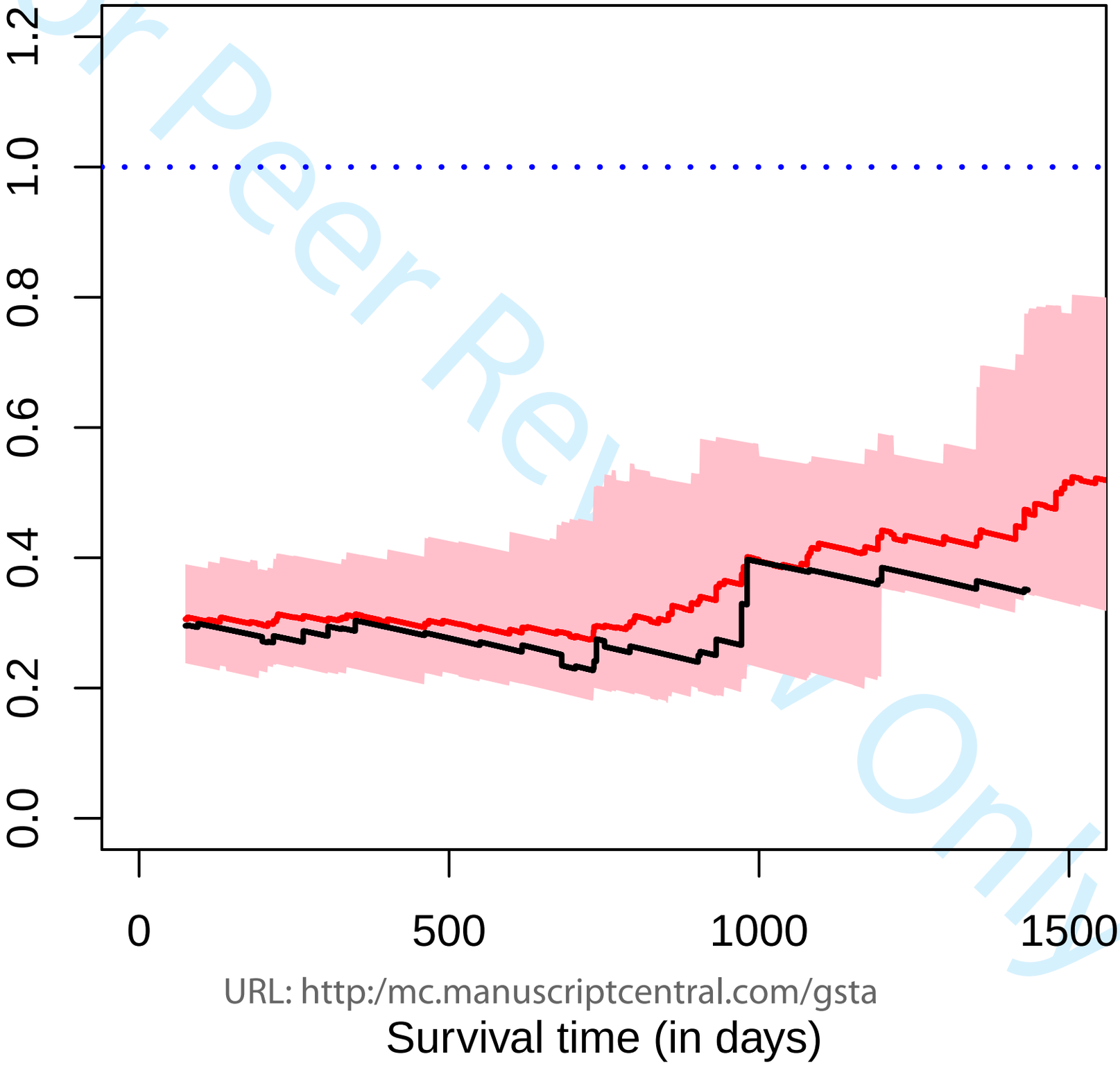


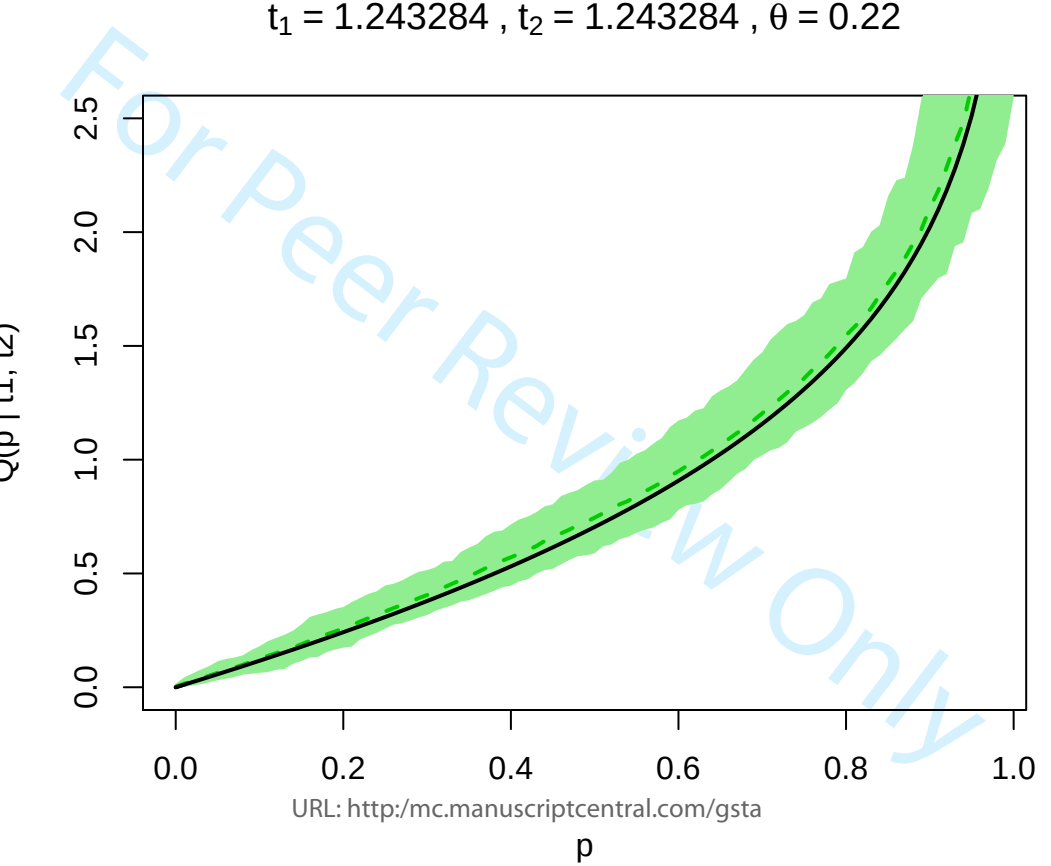


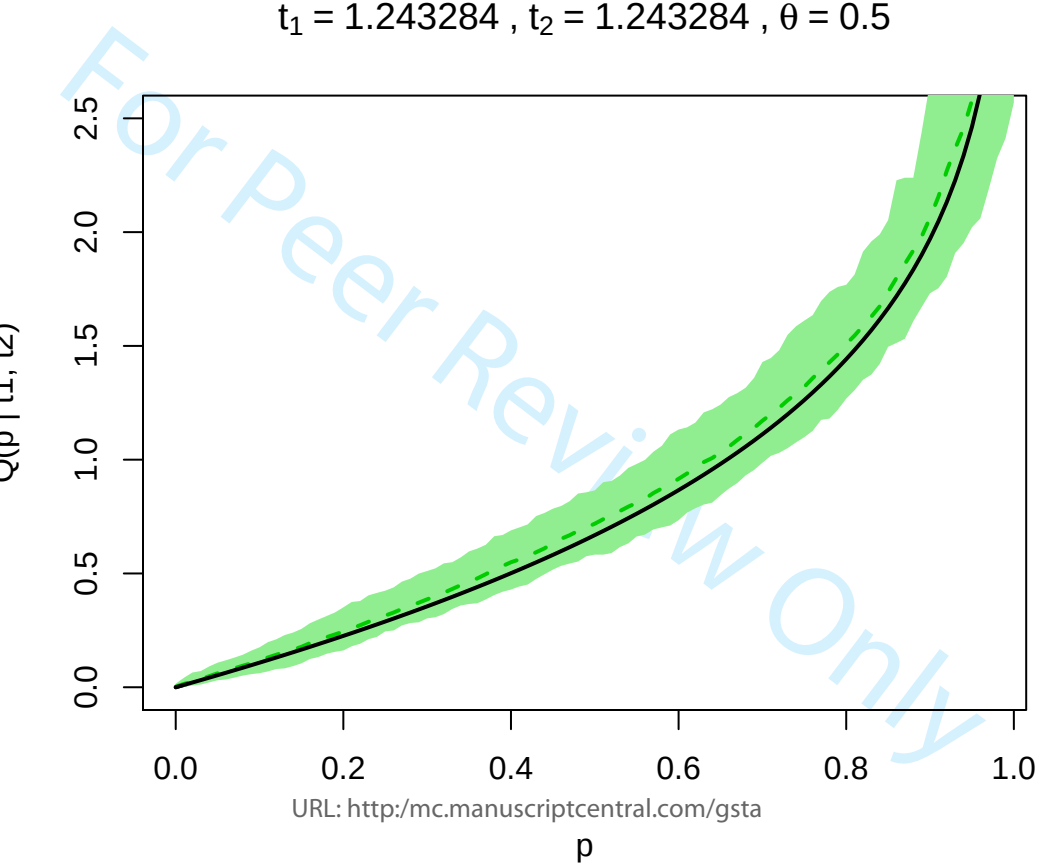




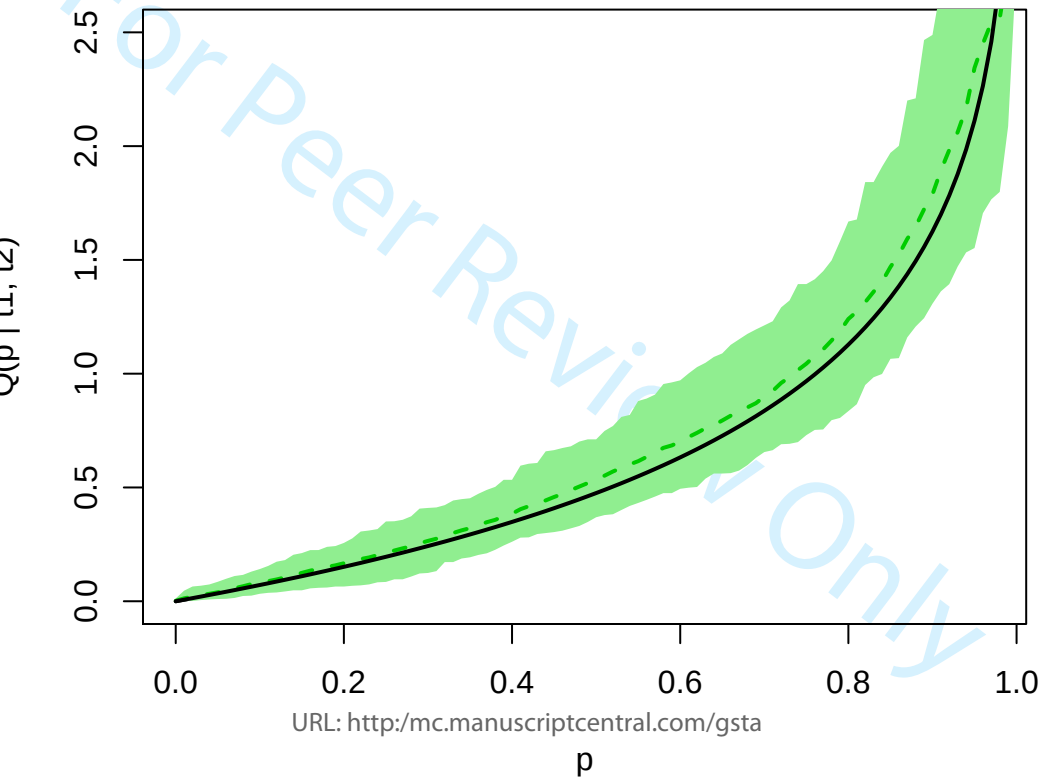




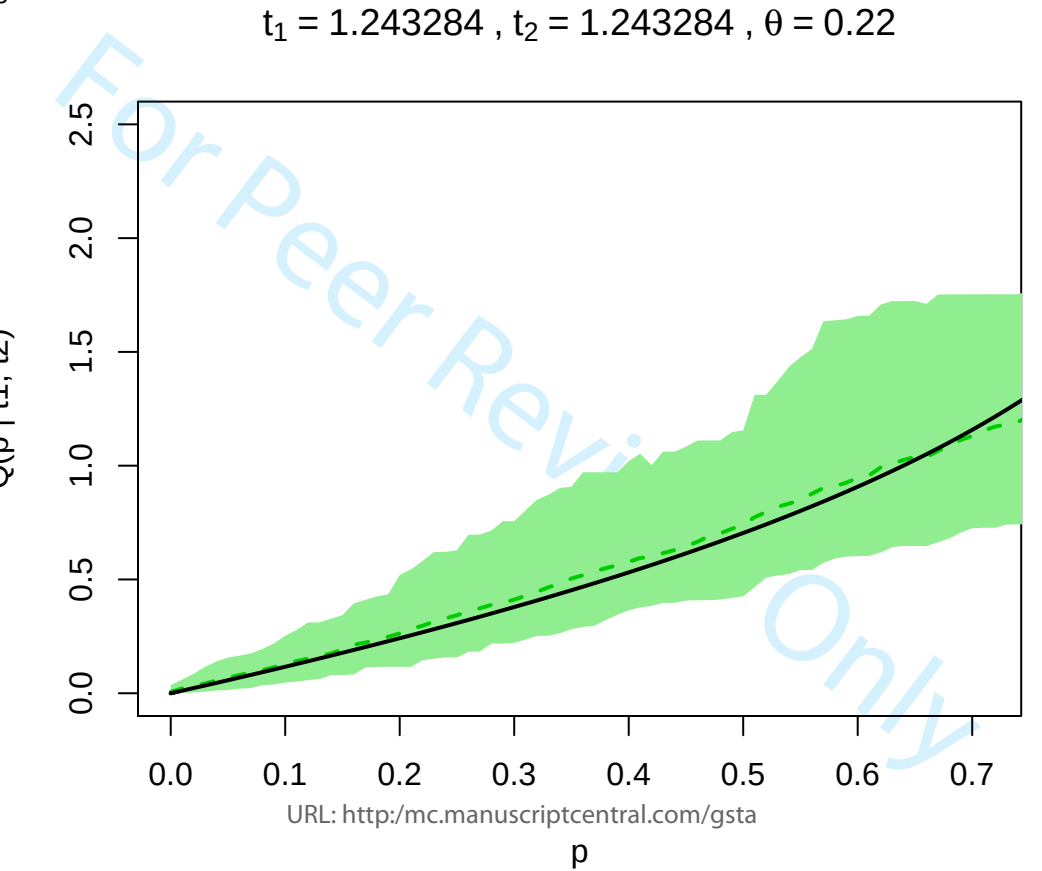


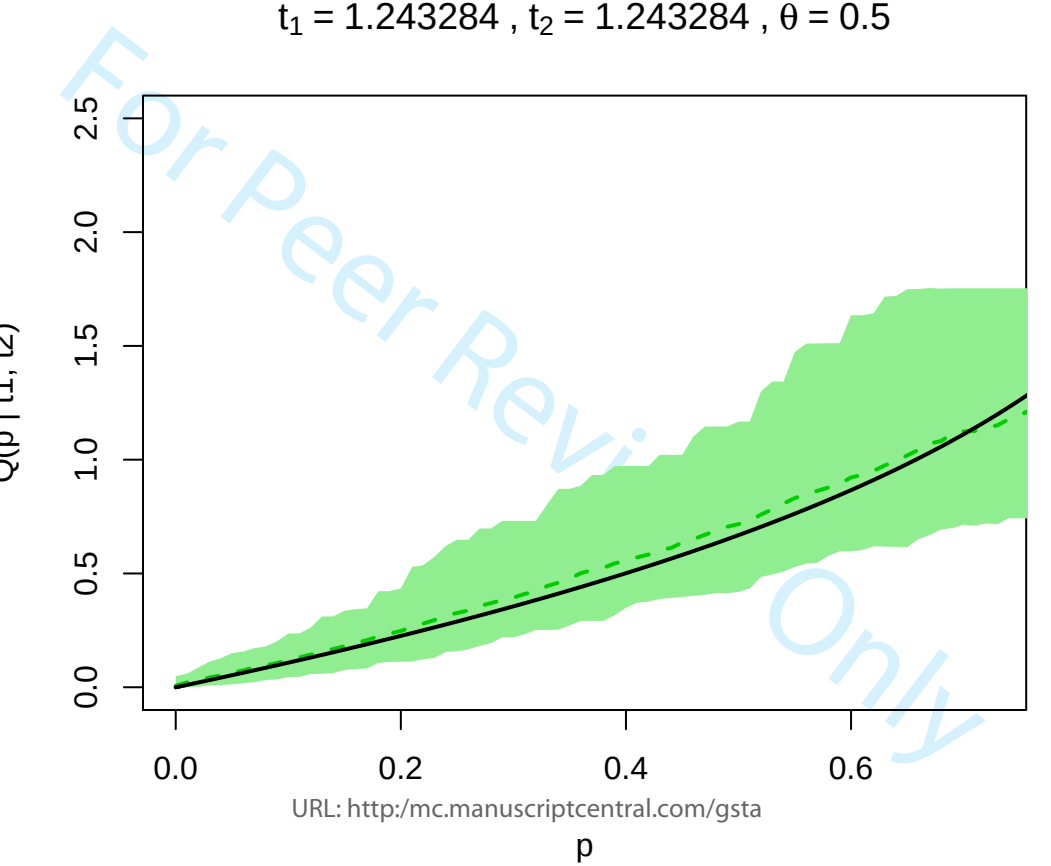


$t_1 = 1.243284$, $t_2 = 1.243284$, $\theta = 2$

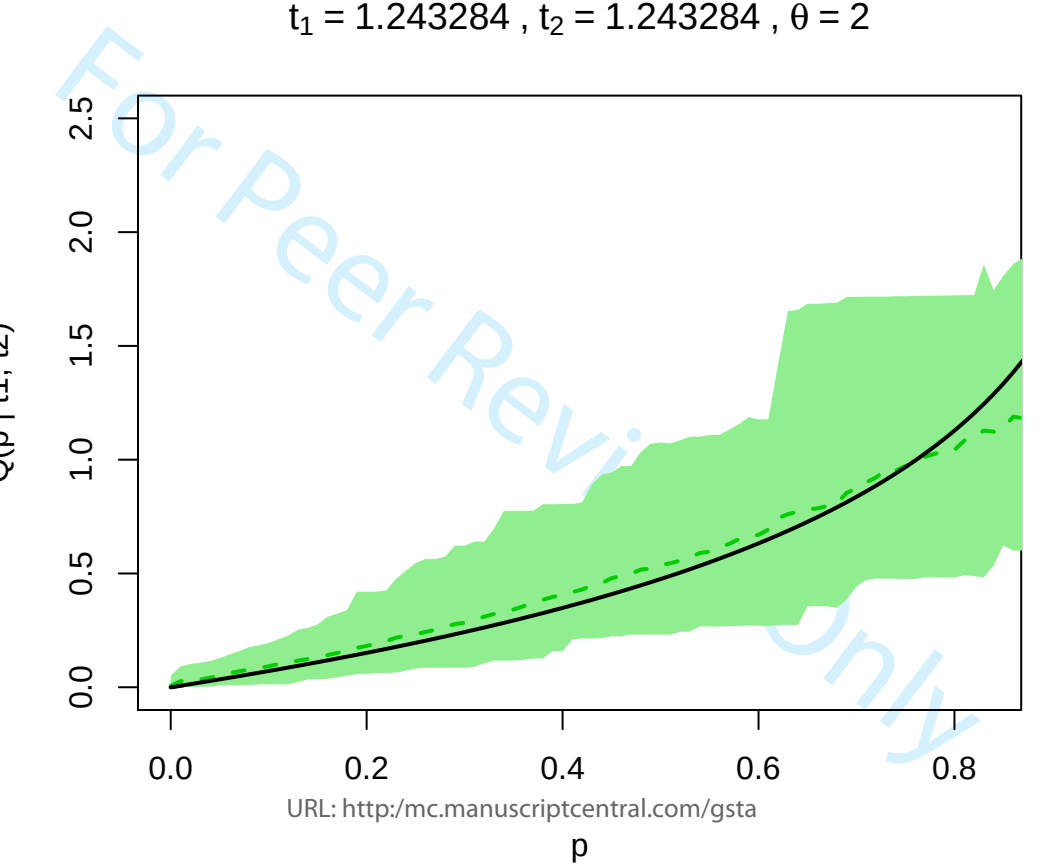


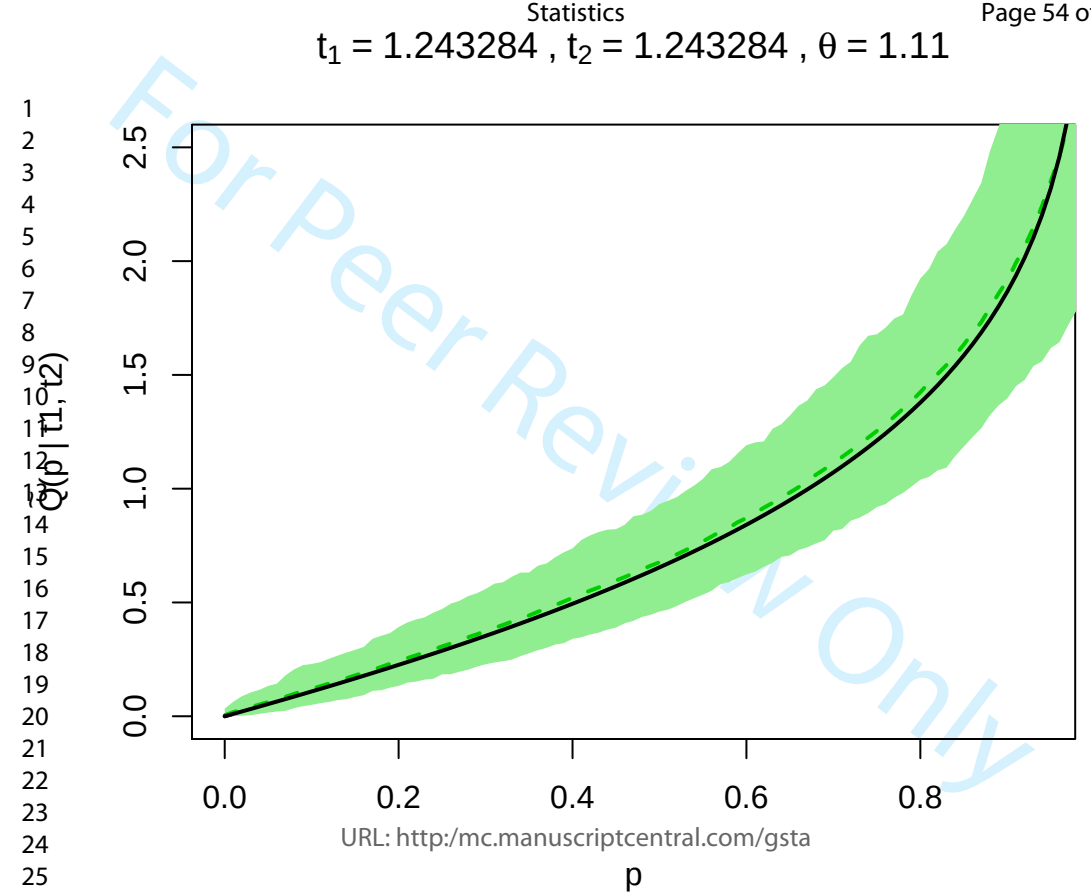
Statistics



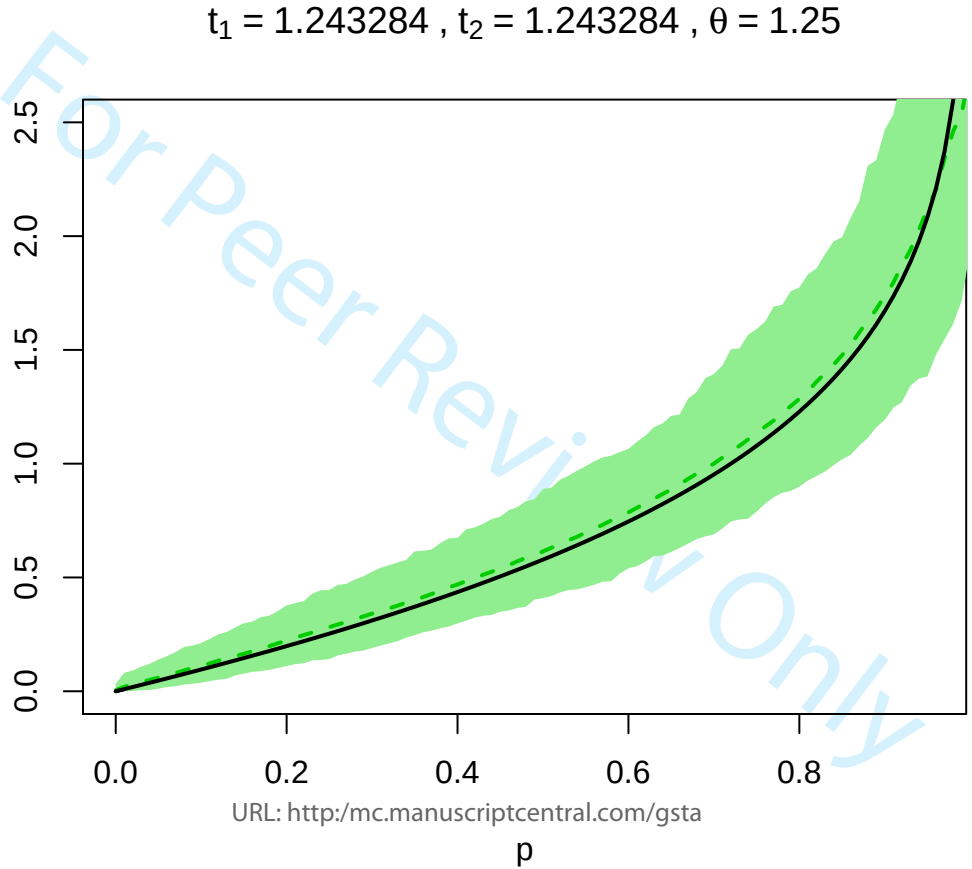


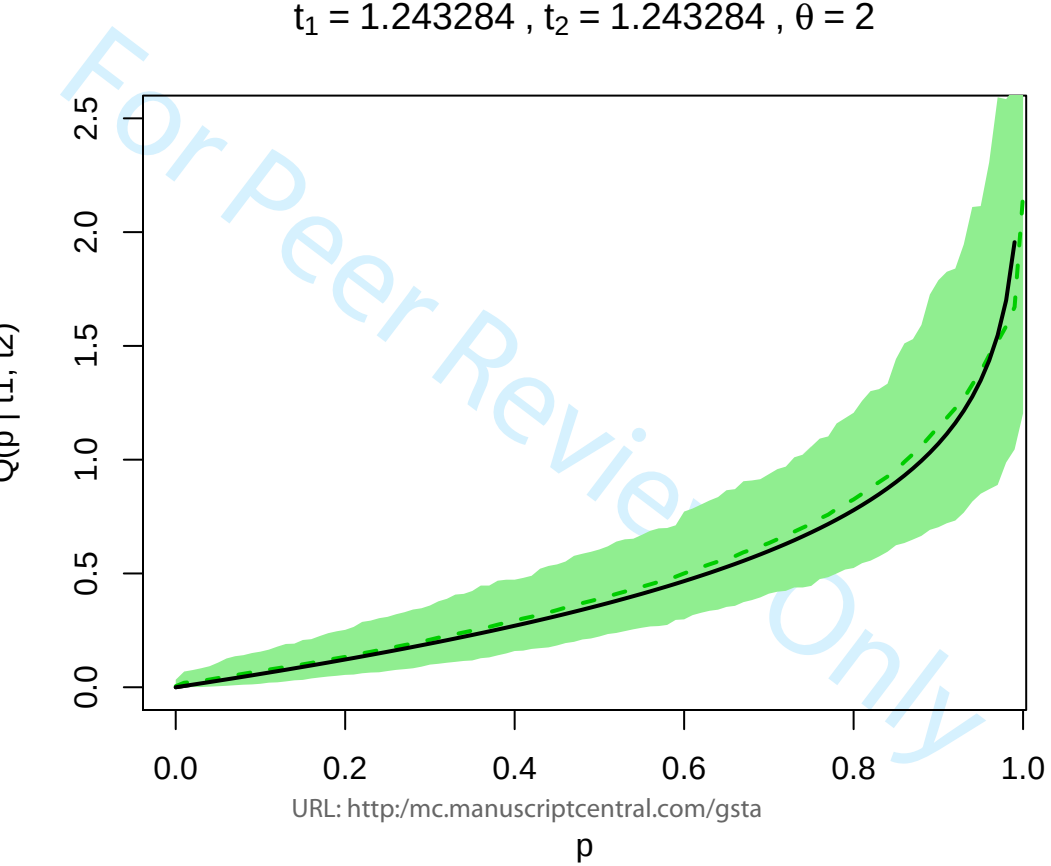
Statistics



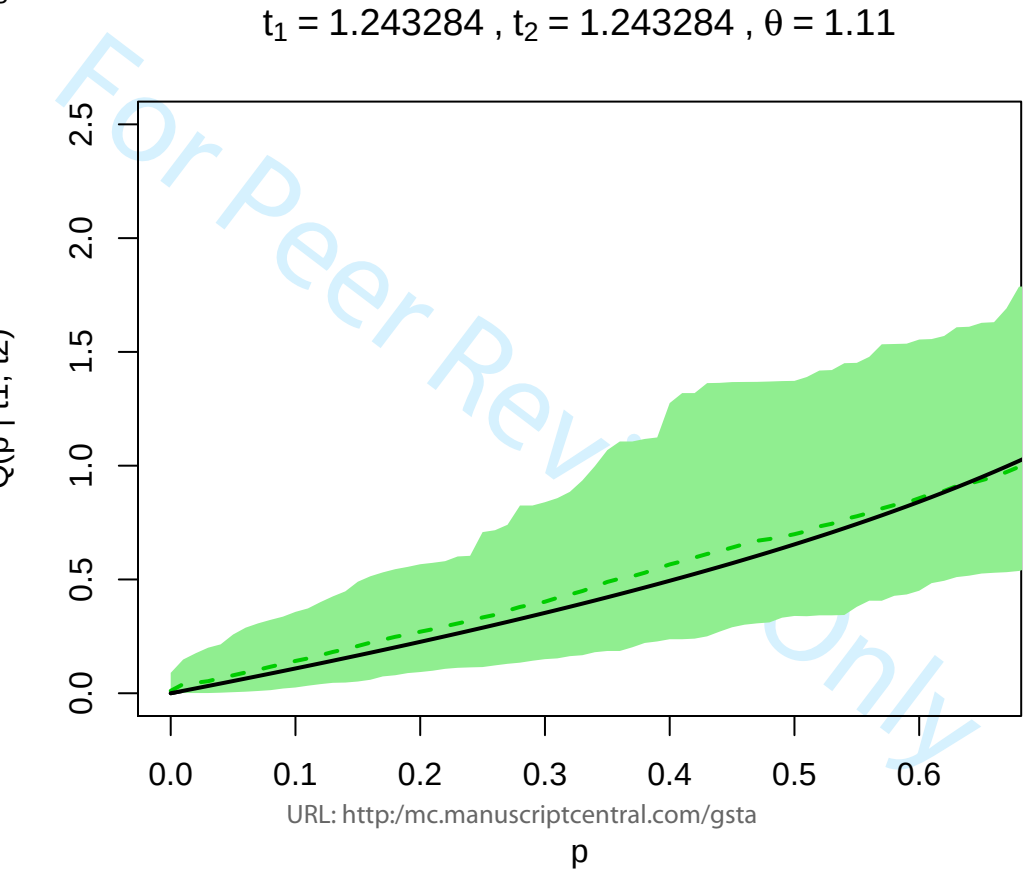


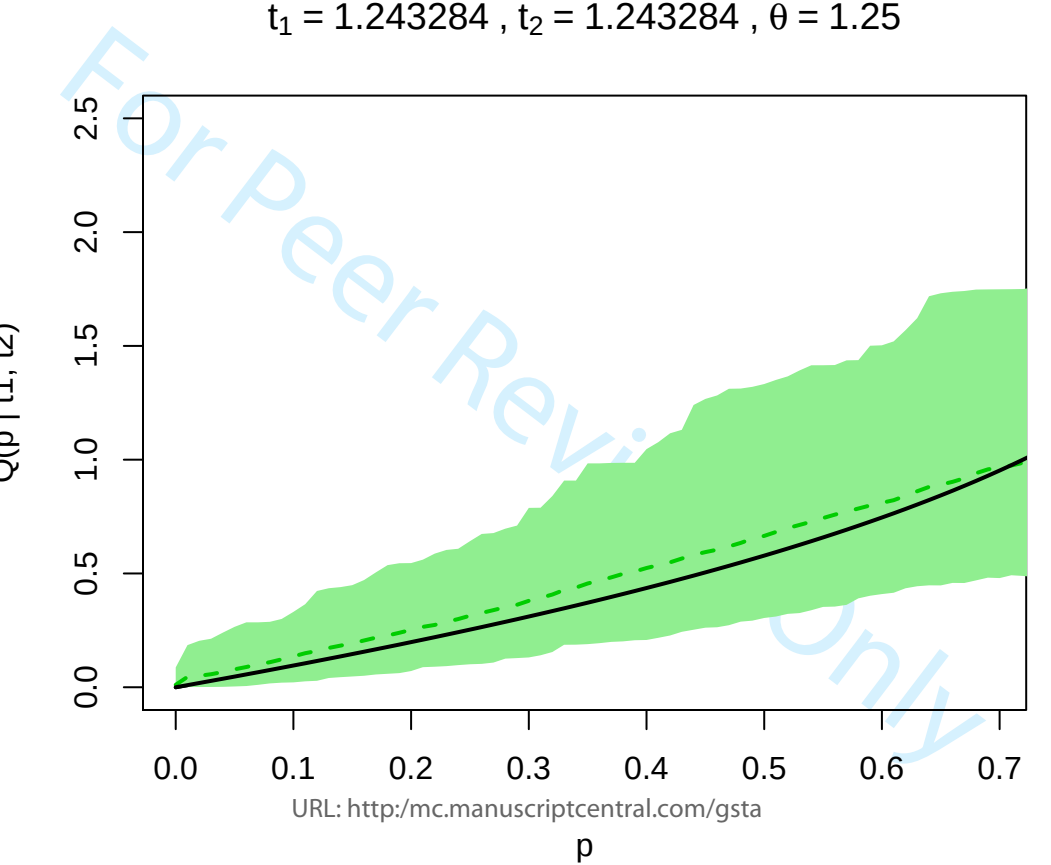
Statistics

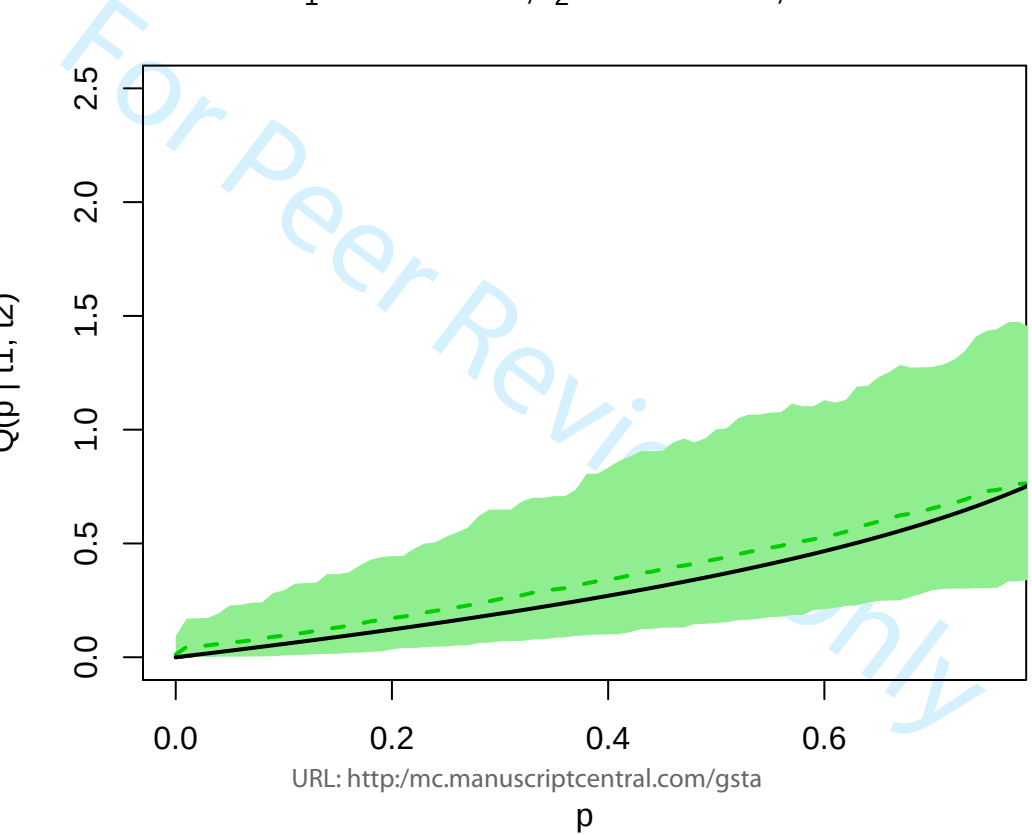




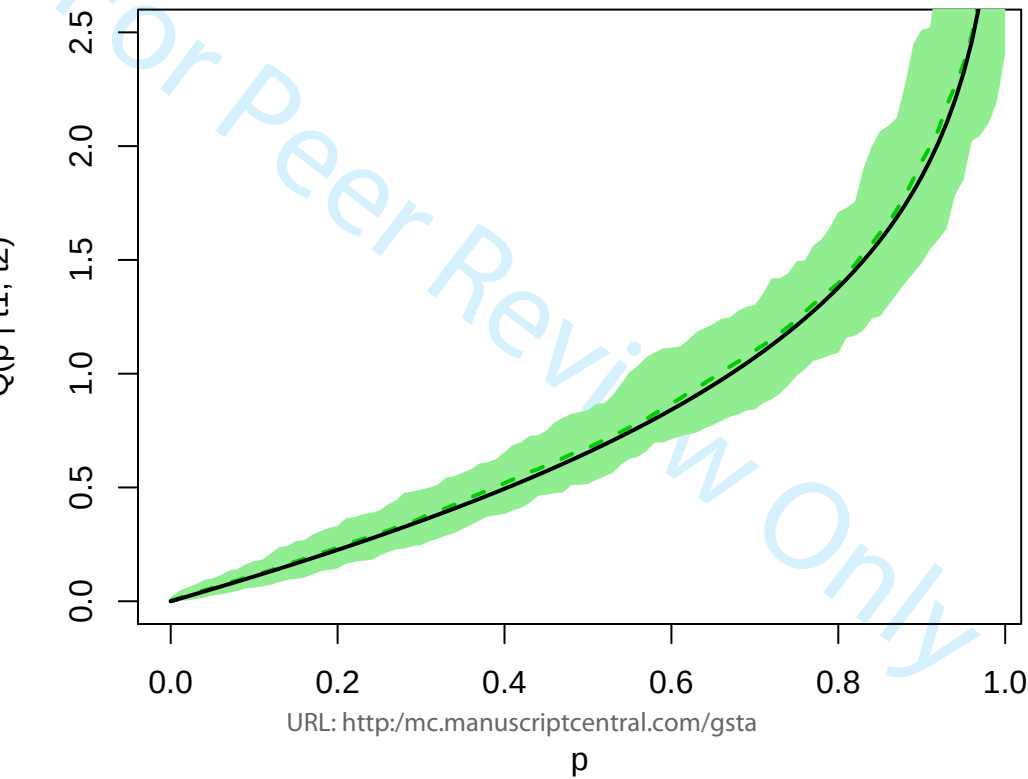
Statistics



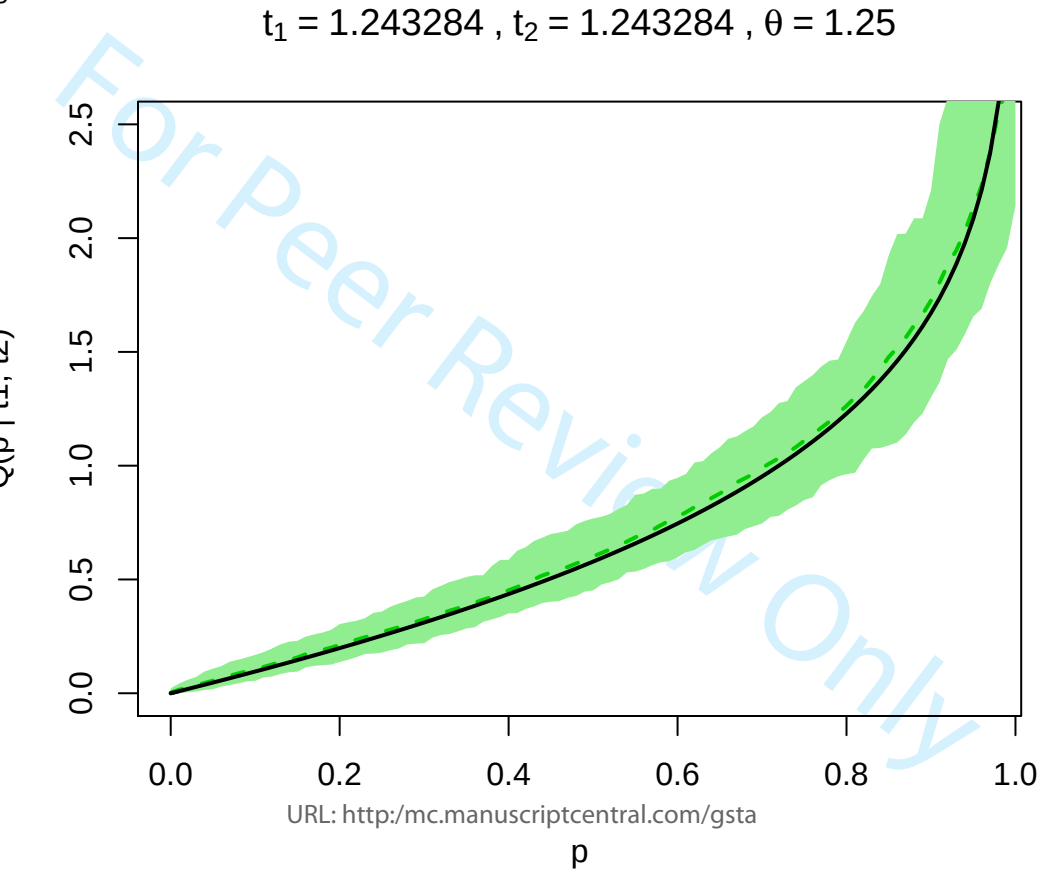


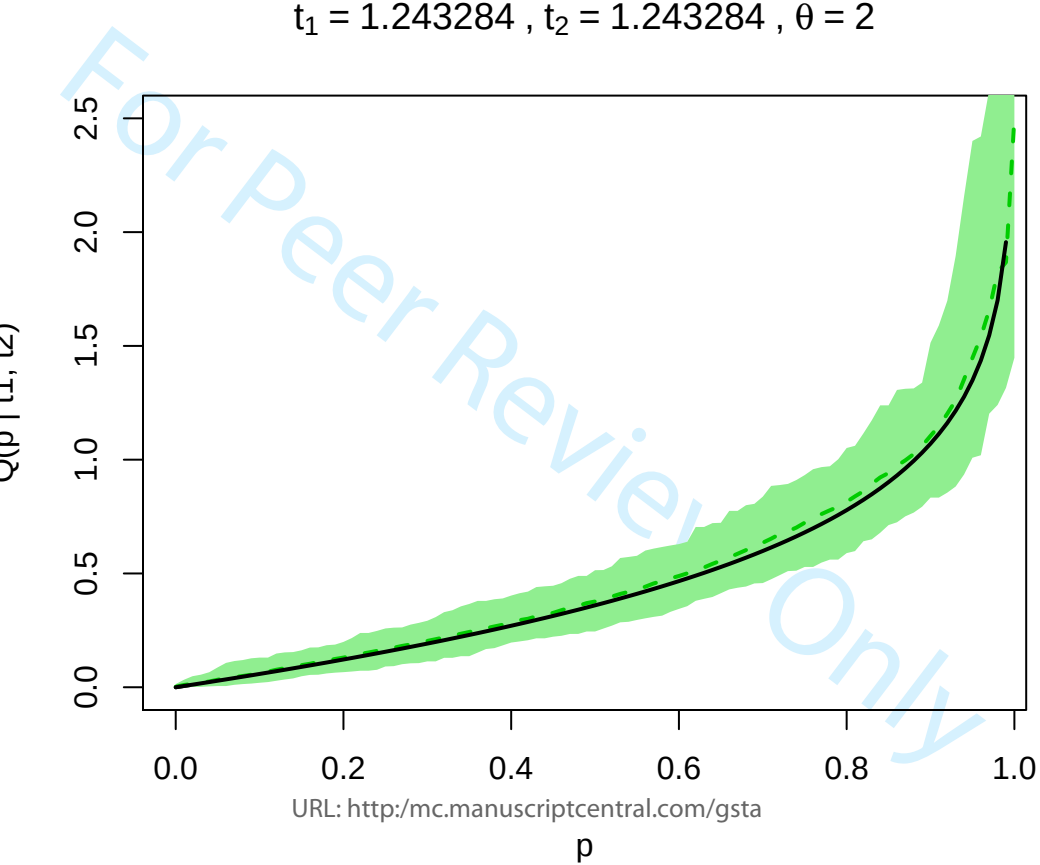


$t_1 = 1.243284$, $t_2 = 1.243284$, $\theta = 1.11$



Statistics





Statistics

$t_1 = 1.243284$, $t_2 = 1.243284$, $\theta = 1.11$