Nonparametric estimation of risk ratios for bivariate data

ABRAMS, Steven; Janssen, P.; SWANEPOEL, Jan & VERAVERBEKE, Noel (2022) Nonparametric estimation of risk ratios for bivariate data. In: JOURNAL OF NONPARAMETRIC STATISTICS, p. 1 -24.

DOI: 10.1080/10485252.2022.2085265
Handle: http://hdl.handle.net/1942/37579

Noname manuscript No.
(will be inserted by the editor)

Nonparametric estimation of risk ratios for bivariate data

**S. Abrams · P. Janssen · J. Swanepoel ·
N. Veraverbeke**

Received: date / Accepted: date

Abstract Inspired by the well-known cross ratio proposed by Clayton, we study a new and appealing alternative risk ratio to describe the relation between the components of a bivariate random vector (T_1, T_2). The new measure is defined as the ratio of the conditional hazard rate function of T_1 at t_1, given that $T_2 \geq t_2$ and the conditional hazard rate function of T_1 at t_1, given that $T_2 < t_2$. A nonparametric estimator is proposed and its asymptotic distribution is obtained via Bernstein-based methods applied to the survival copula of (T_1, T_2) and its partial derivatives. The finite sample performance of the new estimator is studied via simulations. The practical meaning of the risk ratio function and its estimator is illustrated in two real datasets, one on food expenditure and net income and one on the relation between cholesterol and age, and between maximum heart rate achieved and age, for patients suffering from heart disease as compared to control patients without heart disease. Interesting extensions of the proposed risk ratio are given in the discussion section.

Keywords Bernstein estimator · Copulas · Hazard rate · Food expenditure data · Heart disease

S. Abrams
Data Science Institute (DSI), Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), UHasselt, Diepenbeek, Belgium
Global Health Institute (GHI), University of Antwerp, Wilrijk, Belgium
E-mail: steven.abrams@uhasselt.be/Steven.Abrams@uantwerpen.be

P. Janssen
Data Science Institute (DSI), Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), UHasselt, Diepenbeek, Belgium
North-West University, Potchefstroom Campus, Potchefstroom, South Africa

J. Swanepoel
North-West University, Potchefstroom Campus, Potchefstroom, South Africa

N. Veraverbeke
Data Science Institute (DSI), Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), UHasselt, Diepenbeek, Belgium
North-West University, Potchefstroom Campus, Potchefstroom, South Africa

1 Introduction

Bivariate survival distributions, conditional survival distributions and conditional hazard rates are key functions for the statistical analysis of bivariate survival data. They, indeed, provide local information on the relation between the components.

Using copulas it is easy to show that all such quantities can be written in terms of (derivatives) of survival copulas and therefore copulas are the right tool to study conditional distributions and conditional hazard rates in a unified way.

For a bivariate vector of (positive) event times (T_1, T_2), define $S(t_1, t_2) = \mathrm{P}(T_1 > t_1, T_2 > t_2)$ the joint survival function and let, for $j = 1, 2$, $S_j(t_j) = \mathrm{P}(T_j > t_j)$ denote the marginal survival functions with $f_j(t_j)$ the corresponding densities. Copulas provide a natural way to capture the relation between T_1 and T_2. Indeed, assuming absolute continuous marginal survival functions, there is a unique copula $C(\cdot, \cdot)$ satisfying $S(t_1, t_2) = C[S_1(t_1), S_2(t_2)]$ (Sklar, 1959). Further, with $C^{(1)}(\cdot, \cdot)$ and $C^{(2)}(\cdot, \cdot)$ denoting the partial derivatives of $C(\cdot, \cdot)$ with respect to the first and second component and with $c(\cdot, \cdot) = C^{(1,2)}(\cdot, \cdot)$ the copula density, the following conditional distributions and hazards are easily obtained (see, e.g., Section 3 in Janssen et al. (2016) and Abrams et al. (2020)):

$$\begin{aligned}
\mathrm{P}(T_1 > t_1 \mid T_2 = t_2) &= C^{(2)}\left[S_1(t_1), S_2(t_2)\right], \\
\lambda(t_1 \mid T_2 = t_2) &= \frac{c\left[S_1(t_1), S_2(t_2)\right] f_1(t_1)}{C^{(2)}\left[S_1(t_1), S_2(t_2)\right]}, \\
\mathrm{P}(T_1 > t_1 \mid T_2 \geq t_2) &= \frac{C\left[S_1(t_1), S_2(t_2)\right]}{S_2(t_2)}, \\
\lambda(t_1 \mid T_2 \geq t_2) &= \frac{C^{(1)}\left[S_1(t_1), S_2(t_2)\right] f_1(t_1)}{C\left[S_1(t_1), S_2(t_2)\right]}, \\
\mathrm{P}(T_1 > t_1 \mid T_2 < t_2) &= \frac{S_1(t_1) - C\left[S_1(t_1), S_2(t_2)\right]}{1 - S_2(t_2)}, \\
\lambda(t_1 \mid T_2 < t_2) &= \frac{\left\{1 - C^{(1)}\left[S_1(t_1), S_2(t_2)\right]\right\} f_1(t_1)}{S_1(t_1) - C\left[S_1(t_1), S_2(t_2)\right]}.
\end{aligned}$$

Parametric and semi-parametric models have been proposed to estimate these quantities, e.g., Cox and Oakes (1984) proposed the proportional hazards model of the first kind for $\lambda(t_1 \mid T_2 = t_2)$; and Arnold and Kim (1996) proposed the proportional hazards model of the second kind for $\lambda(t_1 \mid T_2 > t_2)$.

Also risk ratios have been studied, e.g., the cross ratio (CR)

$$\mathrm{CR}(t_1, t_2) = \frac{\lambda(t_1 \mid T_2 = t_2)}{\lambda(t_1 \mid T_2 > t_2)},$$

is introduced by Clayton (1978) and used to describe the local relation between the components of the random vector (T_1, T_2). In Abrams et al. (2020) a smooth nonparametric estimator of the cross ratio is studied and a short review, including references, of earlier work (mainly in the context of frailty modeling) is given.

An appealing alternative to the CR is our newly proposed risk ratio (RR)

$$\mathrm{RR}(t_1, t_2) = \frac{\lambda(t_1 \mid T_2 \geq t_2)}{\lambda(t_1 \mid T_2 < t_2)} \tag{1}$$

This ratio compares the instantaneous risk for event one to happen at time $T_1 = t_1$ in the group with $T_2 \geq t_2$ (occurrence of event two later than or equal to time t_2) and the group with $T_2 < t_2$ (occurrence of event two before t_2). Consequently, $\mathrm{RR}(t_1, t_2) = 1$ for all (t_1, t_2) if T_1 and T_2 are independent. Interesting to note is that the conditioning in (1), i.e., $T_2 \geq t_2$ versus $T_2 < t_2$, partitions the whole population in a natural way whereas this partitioning property does not hold for the cross ratio function. In a number of applications the conditioning in (1) might be more appealing than the conditioning used for the cross ratio. For example, if T_2 represents a prognostic index, it is interesting to compare the instantaneous risk for groups with $T_2 \geq t_2$ and $T_2 < t_2$ (where t_2 defines some threshold value for the prognostic index). Moreover, several possible extensions of the risk ratio that emerge from the risk ratio defined by (1), e.g., comparing the two disjoint subpopulations $\{t_{21} \leq T_2 < t_{22}\}$ and $\{t_{23} \leq T_2 < t_{24}\}$, are discussed in Section 6.

In this paper, we propose a nonparametric Bernstein type estimator for the risk ratio and we study the finite sample behaviour and asymptotic distributional behaviour of this estimator. Since copulas are defined on $[0, 1] \times [0, 1]$, the choice for Bernstein type estimators is logical (given the uniform convergence of Bernstein approximations). The reader is referred to Section 6 for a further discussion. We use a food expenditure dataset (Example 1) and a heart disease dataset (Example 2) to demonstrate the practical use of the risk ratio $\mathrm{RR}(t_1, t_2)$. The paper is organized as follows. In Section 2 we propose, for complete data, the nonparametric, Bernstein type, estimator for the risk ratio $\mathrm{RR}(t_1, t_2)$ and we give the risk ratios corresponding to the Clayton, Gumbel and Frank copula. Simulation results are presented in Section 3. The asymptotic distributional behaviour of the Bernstein type estimator of $\mathrm{RR}(t_1, t_2)$ is established in Section 4. The real data examples are discussed in Section 5. In the discussion section (Section 6), we show that, based on the ideas and results in the present paper, challenging new research questions on risk ratios for, e.g., right-censored data emerge and we indicate the methodological hurdles to be solved. Extra details on the simulations are collected in the Supplementary Material. All R code and datasets are available on www.ibiostat.be.

2 Methods and materials

2.1 Nonparametric estimator

The risk ratio in (1) can be written as

$$\mathrm{RR}(t_1, t_2) = \frac{\lambda(t_1 \mid T_2 \geq t_2)}{\lambda(t_1 \mid T_2 < t_2)} = \frac{C^{(1)}[S_1(t_1), S_2(t_2)]}{C[S_1(t_1), S_2(t_2)]} \times \frac{\{S_1(t_1) - C[S_1(t_1), S_2(t_2)]\}}{1 - C^{(1)}[S_1(t_1), S_2(t_2)]} \tag{2}$$

$$= \left\{ \frac{S_1(t_1)}{C[S_1(t_1), S_2(t_2)]} - 1 \right\} \times \left\{ \frac{1}{C^{(1)}[S_1(t_1), S_2(t_2)]} - 1 \right\}^{-1}. \tag{3}$$

Note that $S_1(t_1) > C[S_1(t_1), S_2(t_2)]$ and $C^{(1)}[S_1(t_1), S_2(t_2)] = P(T_2 \geq t_2 \mid T_1 = t_1) \leq 1$ leading to $\mathrm{RR}(t_1, t_2) \geq 0$. To obtain a nonparametric Bernstein type estimator, denoted by $\widehat{\mathrm{RR}}_m(t_1, t_2)$, for the risk ratio $\mathrm{RR}(t_1, t_2)$, we replace in (2) and (3) the quantities $C[S_1(t_1), S_2(t_2)]$, $C^{(1)}[S_1(t_1), S_2(t_2)]$, $S_1(t_1)$ and $S_2(t_2)$ by

$C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]$, $C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]$, $S_{1n}(t_1)$ and $S_{2n}(t_2)$, respectively, where

$$C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] = \sum_{k=0}^{m}\sum_{l=0}^{m} C_n\left(\frac{k}{m}, \frac{l}{m}\right) P_{m,k}[S_{1n}(t_1)] P_{m,l}[S_{2n}(t_2)],$$

$$C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] = m \sum_{k=0}^{m-1}\sum_{l=0}^{m} \left[C_n\left(\frac{k+1}{m}, \frac{l}{m}\right) - C_n\left(\frac{k}{m}, \frac{l}{m}\right)\right] \times P_{m-1,k}[S_{1n}(t_1)] P_{m,l}[S_{2n}(t_2)],$$

with

$$\begin{aligned} P_{m,k}(u) &= \binom{m}{k} u^k (1-u)^{m-k}, \\ C_n(u,v) &= S_n\left[S_{1n}^{-1}(u), S_{2n}^{-1}(v)\right], \\ S_n(t_1,t_2) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_{1i} > t_1, T_{2i} > t_2), \\ S_{1n}(t_1) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_{1i} > t_1) = 1 - F_{1n}(t_1), \\ S_{2n}(t_2) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_{2i} > t_2) = 1 - F_{2n}(t_2), \end{aligned}$$

and m is the so-called Bernstein order. The function $\mathbb{1}(\cdot)$ is the indicator function where $\mathbb{1}(A) = 1$ if condition A holds and zero otherwise.

The nonparametric Bernstein-based estimator for the risk ratio is then:

$$\widehat{\mathrm{RR}}_m(t_1,t_2) = \frac{C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]}{C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]} \times \frac{\{S_{1n}(t_1) - C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]\}}{\left\{1 - C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]\right\}}.$$

To estimate the risk ratio, several approaches can be considered: one can think in terms of conditional hazards, of bivariate survival functions (and its derivatives), or of copulas (and its derivatives). Since copulas capture the relation between T_1 and T_2, the copula approach is a logical choice. Given that copulas are defined on $[0,1] \times [0,1]$ and given the Weierstrass approximation theorem, it is natural to use a truncated series based on Bernstein polynomials as nonparametric estimators for the copula and the copula derivative in the risk ratio expression (3). The advantages of using Bernstein estimators have been described in several papers in the literature: Janssen et al. (2012, 2014, 2016); Abrams et al. (2020); Bouezmarni et al. (2009, 2013). In Section 1 of Janssen et al. (2016), a detailed discussion is given on the superior behaviour of Bernstein estimators. For example, the order of the asymptotic variance is typically $O(m^{1/2}/n)$ in interior points (also in our present Theorem 1) versus $O(m/n)$ for kernel smoothers (Sancetta and Satchell, 2004; Leblanc, 2012).

2.2 Computational formulas

For the random sample $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ let, for $i = 1, 2$, $T_{i(1)} \leq T_{i(2)} \leq \ldots \leq T_{i(n)}$ denote the ordered T_{ij}-values; and let R_{ij} denote the rank of T_{ij}, $j = 1, \ldots, n$. We use the following computational formulas:

$$C_{m,n}(u_1, u_2) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=\left\lfloor \frac{m(n-R_{1j})}{n+1} \right\rfloor +1}^{m} P_{m,k}(u_1) \times \sum_{l=\left\lfloor \frac{m(n-R_{2j})}{n+1} \right\rfloor +1}^{m} P_{m,l}(u_2),$$

$$C^{(1)}_{m,n}(u_1, u_2) = \frac{m}{n} \sum_{j=1}^{n} P_{m-1, \left\lfloor \frac{m(n-R_{1j})}{n+1} \right\rfloor}(u_1) \times \sum_{l=\left\lfloor \frac{m(n-R_{2j})}{n+1} \right\rfloor +1}^{m} P_{m,l}(u_2),$$

where $\lfloor \cdot \rfloor$ refers to the floor notation; i.e., $\lfloor 2.3 \rfloor = 2$.

2.3 Risk ratio for Archimedean copulas

2.3.1 Clayton copula

First, we consider the Clayton copula ($\theta \in [-1, \infty) \setminus \{0\}$) given by

$$C_\theta(u, v) = \left\{ \max \left[u^{-\theta} + v^{-\theta} - 1, 0 \right] \right\}^{-1/\theta}$$

Hence, the first derivative with respect to u is

$$C^{(1)}_\theta(u, v) = u^{-(\theta+1)} C^{\theta+1}_\theta(u, v),$$

for $C(u, v) > 0$, which is fulfilled if $\theta > 0$ and implies that $v > (1 - u^{-\theta})^{-1/\theta}$ if $\theta < 0$ (Nelsen, 2006). In order to avoid restrictions on the domain of the risk ratio, we will restrict attention to positive parameter values for θ in the simulation study (Section 3).

Consequently, the aforementioned risk ratio is:

$$\begin{aligned} \mathrm{RR}(t_1, t_2) &= \left\{ \frac{S_1(t_1)}{C_\theta[S_1(t_1), S_2(t_2)]} - 1 \right\} \times \left\{ \frac{S_1^{\theta+1}(t_1)}{C_\theta^{\theta+1}[S_1(t_1), S_2(t_2)]} - 1 \right\}^{-1} \\ &= \frac{\psi_c(t_1, t_2) - 1}{\psi_c^{\theta+1}(t_1, t_2) - 1}, \end{aligned}$$

where

$$\psi_c(t_1, t_2) = \frac{S_1(t_1)}{C_\theta[S_1(t_1), S_2(t_2)]}.$$

If $\theta \to 0$, the risk ratio tends to one (implying independence between the event times). Furthermore, one can easily show that $\mathrm{RR}(t_1, t_2) < 1$ for $\theta > 0$, and $\mathrm{RR}(t_1, t_2) > 1$ for $\theta \in [-1, 0)$. For exponentially distributed event times T_i with intensities λ_i, $i = 1, 2$, the risk ratio is given by

$$\mathrm{RR}(t_1, t_2) = \frac{\exp(-\lambda_1 t_1) \left[\exp(\theta \lambda_1 t_1) + \exp(\theta \lambda_2 t_2) - 1 \right]^{1/\theta} - 1}{\exp[-\lambda_1(\theta + 1) t_1] \left[\exp(\theta \lambda_1 t_1) + \exp(\theta \lambda_2 t_2) - 1 \right]^{(\theta+1)/\theta} - 1}.$$

In Fig. 1, we graphically depict the functional form of the risk ratio for varying values of the θ parameters (i.e., $\theta = -1/3, -0.2, 0, 0.5, 2$) and exponential marginal distributions with unit mean and variance for the event times T_1 and T_2. More specifically, $S_1(t_1) = \exp(-t_1)$ and $S_2(t_2) = \exp(-t_2)$ represent the marginal survival functions for T_1 and T_2, respectively, and we choose t_2 to be equal to the solutions of the equations $S_2(t_2) = 0.25, 0.5$ or 0.75, i.e., the first quartile, median event time and third quartile of the distribution of T_2 being $t_2 = -\ln(0.75) = 0.2877$, $t_2 = -\ln(0.5) = 0.6931$ and $t_2 = -\ln(0.25) = 1.3863$, respectively. Furthermore, we also show the risk ratio curve in case $t_2 = t_1$ (right lower panel). Note that $\theta = 0$ (corresponding to independence of T_1 and T_2) represents the limiting value when $\theta \to 0$. For this specific example, the range of t_1-values on the x-axes of the plots is chosen to cover 99.5% of the probability mass (i.e., $S_1^{-1}(1 - 0.995) \approx 5.30$). As mentioned previously, if $\theta < 0$, the domain of the risk ratio $\text{RR}(t_1, t_2)$ is constrained in the sense that $t_1 < (1/\theta)\ln\left[1 - \exp(\theta t_2)\right]$ (red lines) (see also Table 1). For the risk ratio on the main diagonal, it should hold that $t_1 = t_2 < (1/\theta)\ln(0.5)$.

[Table 1 about here.]

Kendall's τ is $\theta/(\theta + 2)$ (Nelsen, 2006), implying that the selected values for θ correspond to Kendall's τ values of -0.2, -0.11, 0, 0.2 and 0.5.

[Fig. 1 about here.]

2.3.2 Gumbel copula

Second, a Gumbel copula is considered to induce time-varying association among the infection times, i.e., for $\theta \in [1, \infty)$:

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

with Kendall's τ being equal to $(\theta - 1)/\theta$. We have for the Gumbel copula:

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

The risk ratio has the following functional form:

$$\begin{aligned}\text{RR}(t_1, t_2) = &\left\{\frac{S_1(t_1)}{C_\theta[S_1(t_1), S_2(t_2)]} - 1\right\} \times \\ &\left\{\frac{S_1(t_1)\left\{-\ln\left[S_1(t_1)\right]\right\}^{1-\theta}}{C_\theta[S_1(t_1), S_2(t_2)]\left(-\ln\left\{C_\theta[S_1(t_1), S_2(t_2)]\right\}\right)^{1-\theta}} - 1\right\}^{-1}.\end{aligned}$$

In case of exponential event times with intensities λ_1 and λ_2 for T_1 and T_2, the expression becomes:

$$\begin{aligned}\mathrm{RR}(t_1,t_2) &= \left\{\frac{S_1(t_1)}{C_\theta[S_1(t_1),S_2(t_2)]}-1\right\}\times\\ &\quad\left\{\frac{\exp(-\lambda_1 t_1)(\lambda_1 t_1)^{1-\theta}}{C_\theta[S_1(t_1),S_2(t_2)]\left[(\lambda_1 t_1)^\theta+(\lambda_2 t_2)^\theta\right]^{1/\theta-1}}-1\right\}^{-1}\\ &= \left\{\frac{\exp(-\lambda_1 t_1)}{C_\theta[S_1(t_1),S_2(t_2)]}-1\right\}\left\{\frac{\zeta(t_1,t_2)}{C_\theta[S_1(t_1),S_2(t_2)]}-1\right\}^{-1},\end{aligned}$$

where

$$\zeta(t_1,t_2)=\exp(-\lambda_1 t_1)(\lambda_1 t_1)^{1-\theta}\left[(\lambda_1 t_1)^\theta+(\lambda_2 t_2)^\theta\right]^{1-1/\theta}.$$

For $\theta=1$, $\zeta(t_1,t_2)=\exp(-\lambda_1 t_1)=S_1(t_1)$ and $\mathrm{RR}(t_1,t_2)\equiv 1$. Furthermore, for $\theta>1$, $\mathrm{RR}(t_1,t_2)<1$, for all t_1,t_2. In Fig. 2, we show the risk ratio for $\theta=1$, 10/8 and 2, and exponential marginal distributions with unit mean and variance for the event times T_1 and T_2. Again, we select $t_2=0.2877$, $t_2=0.6931$, $t_2=1.3863$ and $t_2=t_1$. Kendall's $\tau=(\theta-1)/\theta$ is 0, 0.2 and 0.5 for the selected θ values.

[Fig. 2 about here.]

2.3.3 Frank copula

Finally, a Frank copula is considered with parameter θ. The copula is defined as

$$C_\theta(u,v)=-\theta^{-1}\ln\left\{1+\frac{[\exp(-\theta u)-1]\,[\exp(-\theta v)-1]}{\exp(-\theta)-1}\right\},$$

for $\theta\in(-\infty,\infty)\setminus\{0\}$. Consequently, the first derivative of the copula with respect to the first argument is given by:

$$C_\theta^{(1)}(u,v)=\frac{\exp(-\theta u)\,[\exp(-\theta v)-1]}{[\exp(-\theta)-1]+[\exp(-\theta u)-1]\,[\exp(-\theta v)-1]}.$$

Hence,

$$\begin{aligned}\left\{\frac{1}{C_\theta^{(1)}[S_1(t_1),S_2(t_2)]}-1\right\}^{-1} &= \left\{\frac{C_\theta^{(1)}[S_1(t_1),S_2(t_2)]}{1-C_\theta^{(1)}[S_1(t_1),S_2(t_2)]}\right\}\\ &= \left(\frac{\exp[-\theta S_1(t_1)]\,\{\exp[-\theta S_2(t_2)]-1\}}{\exp(-\theta)-\exp[-\theta S_2(t_2)]}\right),\end{aligned}$$

since

$$1-C_\theta^{(1)}(u,v)=\frac{\exp(-\theta)-\exp(-\theta v)}{[\exp(-\theta)-1]+[\exp(-\theta u)-1]\,[\exp(-\theta v)-1]}.$$

The risk ratio has the following expression:

$$\mathrm{RR}(t_1,t_2) = \left\{\frac{S_1(t_1)}{C_\theta[S_1(t_1),S_2(t_2)]} - 1\right\}\left(\frac{\exp[-\theta S_1(t_1)]\,\{\exp[-\theta S_2(t_2)]-1\}}{\exp(-\theta)-\exp[-\theta S_2(t_2)]}\right).$$

When $\theta \to 0$, the risk ratio $\mathrm{RR}(t_1,t_2)$ tends to one. Furthermore, for $\theta < 0$ (> 0), $\mathrm{RR}(t_1,t_2) > 1$ (< 1), for all t_1, t_2. Fig. 3 presents the risk ratio curves for $\theta = -1.8609$, $10^{-6}(\approx 0)$, 1.8609 and 5.7363, and exponential marginal distributions with unit mean and variance for the event times T_1 and T_2 as before. Again, we select $t_2 = 0.2877$, $t_2 = 0.6931$, $t_2 = 1.3863$ and $t_2 = t_1$. Kendall's $\tau = 1 - 4\theta^{-1}[1 - D_1(\theta)]$, with $D_1(.)$ the Debye function of the first kind defined as

$$D_1(\theta) = \theta^{-1}\int_0^\theta \frac{t}{\exp(t)-1},$$

has values of -0.2, 0, 0.2 and 0.5 for the selected θ values.

[Fig. 3 about here.]

Some further insight in the limiting behaviour of the risk ratio (for the situations in Fig. 1– Fig. 3) can be gained from the discussion in Appendix A of the Supplementary Material.

3 Simulations

In this section, we use simulation results to study the behaviour of the proposed nonparametric estimator for the risk ratio $\mathrm{RR}(t_1,t_2)$. In order to get a closer look at the finite-sample performance of our estimator, denoted by $\widehat{\mathrm{RR}}_m(t_1,t_2)$, for the risk ratio $\mathrm{RR}(t_1,t_2)$, we show $\widehat{\mathrm{RR}}_m(t_1,t_2)$ evaluated in a set of points (t_1,t_2) in the rectangle $[a_1,b_1]\times[a_2,b_2]$, where a_1, a_2, b_1 and b_2 are defined below. Furthermore, we compare the integrated squared error, denoted by ISE, which is approximated by evaluating $\widehat{\mathrm{RR}}_m(t_1,t_2)$ at different points (t_1,t_2) in $[a_1,b_1]\times[a_2,b_2]$. More specifically, we consider points $(t_1,t_2) = (F_1^{-1}(u_1), F_2^{-1}(u_2))$, where (u_1,u_2) are inner grid points in the unit square $[0,1]\times[0,1]$, such that

$$\begin{aligned}ISE_{\mathrm{RR}} &= \int_{a_1}^{b_1}\int_{a_2}^{b_2}\left[\widehat{\mathrm{RR}}_m(t_1,t_2) - \mathrm{RR}(t_1,t_2)\right]^2 \mathrm{d}t_1\mathrm{d}t_2\\ &= \int_{a_1^*}^{b_1^*}\int_{a_2^*}^{b_2^*}\left\{\widehat{\mathrm{RR}}_m\left[F_1^{-1}(u_1),F_2^{-1}(u_2)\right] - \mathrm{RR}\left[F_1^{-1}(u_1),F_2^{-1}(u_2)\right]\right\}^2 \mathrm{d}F_1^{-1}(u_1)\mathrm{d}F_2^{-1}(u_2).\end{aligned}$$

ISE_{RR} is approximated on a bivariate grid of $N_1 \times N_2$ values with equally spaced grid points (i.e., $N_1 = N_2 = N$) in $[a_1^*,b_1^*]\times[a_2^*,b_2^*] \equiv [F_1(a_1),F_1(b_1)]\times[F_2(a_2),F_2(b_2)]$, i.e.,

$$I_{\mathrm{RR}} = \Delta_1\Delta_2\sum_{k=1}^{N_1}\sum_{l=1}^{N_2} w_{kl}\left\{\widehat{\mathrm{RR}}_m\left[F_1^{-1}(u_{1[k]}),F_2^{-1}(u_{2[l]})\right] - \mathrm{RR}\left[F_1^{-1}(u_{1[k]}),F_2^{-1}(u_{2[l]})\right]\right\}^2,$$

where $\Delta_1 = (b_1^* - a_1^*)/N_1$, $\Delta_2 = (b_2^* - a_2^*)/N_2$, $u_{1[k]} = a_1^* + (b_1^* - a_1^*)(k-1)/(N_1-1)$ and $u_{2[l]} = a_2^* + (b_2^* - a_2^*)(l-1)/(N_2-1)$, $k = 1,\ldots,N_1$, $l = 1,\ldots,N_2$. The weights are equal to $w_{kl} = (\mathrm{d}F_1^{-1}(u_1)/\mathrm{d}u_1)(\mathrm{d}F_2^{-1}(u_2)/\mathrm{d}u_2)$ evaluated in $(u_{1[k]},u_{2[l]})$. The mean integrated squared error $MISE_{\mathrm{RR}}$ is then approximated by averaging over $M = 100$ replications based on simulated datasets of sample size n (denoted by MI_{RR}):

$$MI_{\mathrm{RR}} = \frac{1}{M} \sum_{r=1}^{M} I_{\mathrm{RR}}^{(r)},$$

where $I_{\mathrm{RR}}^{(r)}$ is the approximation of ISE based on the r-th simulated dataset. Finally, we also calculate an approximation of the integrated squared bias and integrated variance, defined as:

$$ISB_{\mathrm{RR}} = \mathit{\Delta}_1 \mathit{\Delta}_2 \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} w_{kl} \left\{ \overline{\widehat{\mathrm{RR}}}_m \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right] - \mathrm{RR} \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right] \right\}^2,$$

$$IV_{\mathrm{RR}} = \frac{1}{M} \sum_{r=1}^{M} \left(\mathit{\Delta}_1 \mathit{\Delta}_2 \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} w_{kl} \left\{ \widehat{\mathrm{RR}}_m^{(r)} \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right] - \overline{\widehat{\mathrm{RR}}}_m \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right] \right\}^2 \right)$$

where

$$\overline{\widehat{\mathrm{RR}}}_m \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right] = \frac{1}{M} \sum_{r=1}^{M} \widehat{\mathrm{RR}}_m^{(r)} \left[F_1^{-1}(u_{1[k]}), F_2^{-1}(u_{2[l]}) \right],$$

and $\widehat{\mathrm{RR}}_m^{(r)}$ represents the Bernstein-based estimator for the risk ratio relying on the r-th simulated dataset. Note that the aforementioned quantities add up to the mean integrated squared error. Throughout this simulation study, we select $[a_1^*, b_1^*] = [a_2^*, b_2^*] = [0.1, 0.9]$ and $N_1 = N_2 = 80$ (i.e., implying $\mathit{\Delta}_1 = \mathit{\Delta}_2 = 0.01$). As mentioned in Appendix A of the Supplementary Material, we focus on the interior points of the unit square to avoid boundary issues (see also Appendix B). Confidence limits provided in this section are pointwise simulation-based ones.

We generate n pairs of uncensored event times (t_{1j}, t_{2j}), $j = 1, \ldots, n$ using the 'copula' package in R version 3.3.2. More specifically, random samples (u_{1j}, u_{2j}) are drawn from three different copulas (i.e., Clayton, Gumbel and Frank copulas) after which dependent exponential event times are obtained with constant hazards $\lambda_1(t_1) \equiv \lambda_1 = 1$ and $\lambda_2(t_2) \equiv \lambda_2 = 1$, as follows:

$$t_{ij} = -\frac{\ln(1 - u_{ij})}{\lambda_i} = -\ln(1 - u_{ij}).$$

Since nonparametric Bernstein estimators critically depend on the choice of the Bernstein order m, we explore different choices thereof. More specifically, we select m from a grid of values $A_n = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$.

In this section we look at the Clayton copula. Similar results for the Gumbel and Frank copula are presented in Appendix C in the Supplementary Material. More specifically, we consider the Clayton copula with parameter values for $\theta = 0$ (independence), 0.5 and 2.0. We restrict attention to positive parameter values to avoid boundary constraints (see the discussion related to Table 1). In Table 2, we show MI_{RR}, ISB_{RR} and IV_{RR} for different choices of the Bernstein order m. In Fig. 4, we graphically show a heatplot of the relative difference between the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ averaged over the $M = 100$ replications and the true risk ratio $\mathrm{RR}(t_1, t_2)$ (left panel), i.e.,

$$\frac{\left[(1/M)\sum_{r=1}^{M}\widehat{\mathrm{RR}}_m^{(r)}(t_1, t_2)\right] - \mathrm{RR}(t_1, t_2)}{\mathrm{RR}(t_1, t_2)}.$$

Furthermore, we give the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ (black solid line in the right panel) as a function of t_1 with $t_2 = F_2^{-1}(0.5)$ fixed (right panel) for $\theta = 0$, $m = 5$ and $n = 500$. Pointwise 95% simulation-based confidence bounds (gray dashed lines) and the true risk ratio (red dash-dotted line) are included as well. The shaded colors from dark to light indicate regions bounded by the 2.5th, 5th and 10th percentile (left-hand side) and the 97.5th, 95th and 90th percentile (right-hand side). Contourplots and additional figures are presented in Appendix C in the Supplementary Material.

[Fig. 4 about here.]

In Fig. 5, we graphically show a heatplot of the relative difference between the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ averaged over the M replications and the true risk ratio $\mathrm{RR}(t_1, t_2)$ (left panel). Moreover, the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ is shown (black solid lines in the right panel) as a function of t_1 with $t_2 = F_2^{-1}(0.5)$ fixed for $\theta = 0.5$, $m = 20$ and $n = 500$. Pointwise 95% simulation-based confidence bounds (gray dashed lines) and the true risk ratio (red dashed line) are again included.

Note that the choice $m = 5$ (in Fig. 4) and $m = 20$ (in Fig. 5) is based on the findings in Table 2.

[Table 2 about here.]

[Fig. 5 about here.]

4 Asymptotic normality of the Bernstein risk ratio estimator

Our main theorem gives the asymptotic distributional behaviour of $\widehat{\mathrm{RR}}_m(t_1, t_2) - \mathrm{RR}(t_1, t_2)$, with

$$\widehat{\mathrm{RR}}_m(t_1, t_2) = \frac{C_{m,n}^{(1)}[S_{1n}(t_1), S_{2n}(t_2)]}{C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]} \times \frac{\{S_{1n}(t_1) - C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]\}}{\left\{1 - C_{m,n}^{(1)}[S_{1n}(t_1), S_{2n}(t_2)]\right\}}.$$

Theorem 1 *Assume*

(C1) The copula function $C(.,.)$ has bounded third order partial derivatives on $(0,1)\times(0,1)$,

(C2) $m = Kn^{\alpha}$ with $\frac{2}{5} < \alpha < \frac{3}{5}$, $K > 0$.

Then, for all (t_1, t_2) *such that* $0 < S_1(t_1), S_2(t_2) < 1$ *and* $0 < C^{(1)}[S_1(t_1), S_2(t_2)] < 1$, *as* $n \to \infty$,

$$\left(\frac{n}{m^{1/2}}\right)^{1/2}\left[\widehat{RR}_m(t_1,t_2) - RR(t_1,t_2)\right] \xrightarrow{d} \mathcal{N}\left(0;\, Var_{RR}(t_1,t_2)\right),$$

with

$$Var_{RR}(t_1,t_2) = \frac{RR^2(t_1,t_2)}{2C^{(1)}[S_1(t_1),S_2(t_2)]\left\{1 - C^{(1)}[S_1(t_1),S_2(t_2)]\right\}\sqrt{\pi S_1(t_1)[1-S_1(t_1)]}}\,.$$

The risk ratio RR is a non-linear function of the copula C, the copula derivative $C^{(1)}$ and the marginal survival functions S_1 and S_2. The estimator $\widehat{\mathrm{RR}}_m$ is the same expression with $C_{m,n}$, $C^{(1)}_{m,n}$, S_{1n} and S_{2n}. The first step in the proof is to linearize the difference $\widehat{\mathrm{RR}}_m - RR$ and to find out that the dominating contribution to $\widehat{\mathrm{RR}}_m - RR$ comes from $C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)]$ and that the contributions from $S_{1n} - S_1$ and $C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)]$ are negligible after scaling (Lemma 1). The limiting distribution of the dominating term is then derived in Lemma 2.

Proof We use D, $D_{m,n}$ (D for Derivative); C, $C_{m,n}$ (C for Copula) and S_1, S_{1n} (S for Survival) as shorthand notation for $C^{(1)}[S_1(t_1), S_2(t_2)]$, $C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]$; $C[S_1(t_1), S_2(t_2)]$, $C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)]$ and $S_1(t_1)$, $S_{1n}(t_1)$, respectively.
We then can write

$$\begin{aligned}\widehat{\mathrm{RR}}_m(t_1,t_2) - \mathrm{RR}(t_1,t_2) &= \frac{D_{m,n}\left(S_{1n} - C_{m,n}\right)}{C_{m,n}\left(1 - D_{m,n}\right)} - \frac{D\left(S_1 - C\right)}{C\left(1-D\right)} \\ &= \frac{(1-D)D_{m,n}C(S_{1n} - C_{m,n}) - D(1-D_{m,n})C_{m,n}(S_1 - C)}{(1-D)(1-D_{m,n})CC_{m,n}} \\ &:= \frac{\mathrm{NUM}}{\mathrm{DENOM}}.\end{aligned}$$

The numerator can be rewritten as

$$\begin{aligned}\mathrm{NUM} = {} & C(S_{1n} - C_{m,n})(D_{m,n} - D) - DS_1(C_{m,n} - C)(1 - D_{m,n}) + \\ & D(1 - D_{m,n})C(S_{1n} - S_1).\end{aligned}$$

From the discussion that follows we will obtain that $C_{m,n} - C = o_P(1)$, $D_{m,n} - D = o_P(1)$ and $S_{1n} - S_1 = o_P(1)$. Using a Slutsky argument we therefore have that

$$\begin{aligned}&\widehat{\mathrm{RR}}_m(t_1,t_2) - \mathrm{RR}(t_1,t_2) \\ &\overset{\mathcal{L}}{\sim} \frac{S_1 - C}{(1-D)^2C}(D_{m,n} - D) - \frac{DS_1}{(1-D)C^2}(C_{m,n} - C) + \frac{D}{(1-D)C}(S_{1n} - S_1),\end{aligned}$$

where $\overset{\mathcal{L}}{\sim}$ means that they have the same asymptotic distributional behaviour.
Note that $S_{1n} - S_1 = O_P(n^{-1/2})$. In Lemma 1 we show that $C_{m,n} - C = O_P(n^{-1/2}) + O(m^{-1})$.
Since the scaling factor in our theorem is $n^{1/2}m^{-1/4}$ we have that $n^{1/2}m^{-1/4}(S_{1n} - S_1) = O_P(m^{-1/4}) = o_P(1)$ and $n^{1/2}m^{-1/4}(C_{m,n} - C) = O_P(m^{-1/4}) + O_P(n^{1/2}m^{-5/4}) = o_P(1)$ if $\alpha > 2/5$. The distributional behaviour of $\widehat{\mathrm{RR}}_m(t_1,t_2) - \mathrm{RR}(t_1,t_2)$ is therefore determined by the distributional behaviour of $(D_{m,n} - D)$, as can be seen from Lemma 2. □

Lemma 1 *Assume (C1). Then, with* $0 < S_1(t_1), S_2(t_2) < 1$*, as* $n \to \infty$*,*

$$C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)] = O_P(n^{-1/2}) + O(m^{-1}).$$

Proof We use the following inequalities

$$\begin{aligned}
&\mid C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)] \mid \\
&\leq \mid C_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C[S_{1n}(t_1), S_{2n}(t_2)] \mid \\
&\quad + \mid C[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)] \mid \\
&\leq \sup_{0<u,v<1} \mid C_{m,n}(u,v) - C(u,v) \mid \\
&\quad + \mid C[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)] \mid .
\end{aligned}$$

With

$$B_m(u,v) = \sum_{k=0}^{m} \sum_{l=0}^{m} C\left(\frac{k}{m}, \frac{l}{m}\right) P_{m,k}(u) P_{m,l}(v)$$

we have

$$\begin{aligned}
&\sup_{0<u,v<1} \mid C_{m,n}(u,v) - C(u,v) \mid \\
&\quad \leq \sup_{0<u,v<1} \mid C_{m,n}(u,v) - B_m(u,v) \mid + \sup_{0<u,v<1} \mid B_m(u,v) - C(u,v) \mid \\
&\quad \leq \sup_{0<u,v<1} \mid C_n(u,v) - C(u,v) \mid + \sup_{0<u,v<1} \mid B_m(u,v) - C(u,v) \mid .
\end{aligned}$$

Weak convergence of the process $C_n(u,v) - C(u,v)$ implies $\sup_{0<u,v<1} \mid C_n(u,v) - C(u,v) \mid = O_P(n^{-1/2})$, see Fermanian et al. (2004). Moreover, under (C1), we have $\sup_{0<u,v<1} \mid B_m(u,v) - C(u,v) \mid = O(m^{-1})$ (see (5) in Janssen et al. (2012)).

This yields

$$\sup_{0<u,v<1} \mid C_{m,n}(u,v) - C(u,v) \mid = O_P(n^{-1/2}) + O(m^{-1}).$$

Young's form of the Taylor expansion combined with $S_{jn}(t_j) - S_j(t_j) = O_P(n^{-1/2})$, $j = 1, 2$, gives

$$\mid C[S_{1n}(t_1), S_{2n}(t_2)] - C[S_1(t_1), S_2(t_2)] \mid = O_P(n^{-1/2}). \qquad \square$$

To study the distributional behaviour of

$$C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)],$$

we first introduce the following quantities:

$$\beta^o_m[S_1(t_1), S_2(t_2)] = \sum_{k=0}^{m-1} \sum_{l=0}^{m} C^{(1)}\left(\frac{k}{m-1}, \frac{l}{m}\right) P_{m-1,k}[S_1(t_1)] P_{m,l}[S_2(t_2)],$$

a Bernstein-Weierstrass approximation for $C^{(1)}[S_1(t_1), S_2(t_2)]$, and

$$\beta_m[S_1(t_1), S_2(t_2)] = m \sum_{k=0}^{m-1} \sum_{l=0}^{m} \left[C\left(\frac{k+1}{m}, \frac{l}{m}\right) - C\left(\frac{k}{m}, \frac{l}{m}\right)\right] P_{m-1,k}[S_1(t_1)] P_{m,l}[S_2(t_2)],$$

a first order differential version of $\beta_m^o[S_1(t_1), S_2(t_2)]$.

Now use the following decomposition

$$
\begin{aligned}
&C_{m,n}^{(1)}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)] \\
&= \left(\left\{ C_{m,n}^{(1)}[S_{1n}(t_1), S_{2n}(t_2)] - \beta_m[S_{1n}(t_1), S_{2n}(t_2)] \right\} - \left\{ C_{m,n}^{(1)}[S_1(t_1), S_2(t_2)] - \beta_m[S_1(t_1), S_2(t_2)] \right\} \right) \\
&\quad + \left\{ C_{m,n}^{(1)}[S_1(t_1), S_2(t_2)] - \beta_m^o[S_1(t_1), S_2(t_2)] \right\} \\
&\quad + \left\{ \beta_m[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)] \right\} \\
&\quad + \{ \beta_m^o[S_1(t_1), S_2(t_2)] - \beta_m[S_1(t_1), S_2(t_2)] \} \\
&:= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}
$$

Assuming the conditions of our main theorem, we can apply Theorem 2 in Janssen et al. (2016), i.e., we obtain the following asymptotic behaviour of II:

$$
\begin{aligned}
&\left(\frac{n}{m^{1/2}} \right)^{1/2} \left\{ C_{m,n}^{(1)}[S_1(t_1), S_2(t_2)] - \beta_m^o[S_1(t_1), S_2(t_2)] \right\} \\
&\quad \xrightarrow{d} \mathcal{N}\left(0, \frac{C^{(1)}[S_1(t_1), S_2(t_2)] \left\{ 1 - C^{(1)}[S_1(t_1), S_2(t_2)] \right\}}{2\sqrt{\pi S_1(t_1)[1 - S_1(t_1)]}} \right).
\end{aligned}
$$

Assuming (C1) and $\alpha < 3/5$ we also can apply Lemma 3 in the Supplementary Material of Janssen et al. (2016). This gives

$$
I = O_P\left(m^{13/12} n^{-1} [\ln(n)]^{1/2} \{\ln[\ln(n)]\}^{1/2} \right).
$$

An order bound for III is obtained as follows:

$$
\begin{aligned}
\mathrm{III} &= \{ \beta_m[S_{1n}(t_1), S_{2n}(t_2)] - \beta_m^o[S_{1n}(t_1), S_{2n}(t_2)] \} \\
&\quad + \left\{ \beta_m^o[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_{1n}(t_1), S_{2n}(t_2)] \right\} \\
&\quad + \left\{ C^{(1)}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)] \right\} \\
&= O(m^{-1}) + O_P(n^{-1/2}),
\end{aligned}
$$

where the order relations follow from the detailed discussion on Section 2 of Janssen et al. (2016). That same discussion implies that $\mathrm{IV} = O(m^{-1})$.

Finally note that

$$
\left(\frac{n}{m^{1/2}} \right)^{1/2} I = O_P(m^{5/6} n^{-1/2} [\ln(n)]^{1/2} \{\ln[\ln(n)]\}^{1/2}) = o_P(1),
$$

for $m = Kn^{\alpha}$, $\alpha < 3/5$, and that

$$
\left(\frac{n}{m^{1/2}} \right)^{1/2} (\mathrm{III} + \mathrm{IV}) = O\left(\frac{n^{1/2}}{m^{5/4}} + \frac{1}{m^{1/4}} \right) = o_P(1),
$$

for $m = Kn^{\alpha}$, $\alpha > 2/5$.

Collecting all these results, we have shown the following lemma.

Lemma 2 *Assume (C1) and* $m = Kn^{\alpha}$, $\frac{2}{5} < \alpha < \frac{3}{5}$, $K > 0$. *Then, with* $0 < S_1(t_1), S_2(t_2) < 1$, *as* $n \to \infty$,

$$\left(\frac{n}{m^{1/2}}\right)^{1/2} \left\{C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)]\right\}$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{C^{(1)}[S_1(t_1), S_2(t_2)]\left\{1 - C^{(1)}[S_1(t_1), S_2(t_2)]\right\}}{2\sqrt{\pi S_1(t_1)[1 - S_1(t_1)]}}\right). \qquad \square$$

We now have all the ingredients to prove our main theorem. First note that $(S_1 - C)/[C(1-D)^2]$, the coefficient of $D_{m,n} - D$ in the decomposition of $\widehat{\mathrm{RR}}_m(t_1, t_2) - \mathrm{RR}(t_1, t_2)$ can be rewritten as

$$\frac{\mathrm{RR}(t_1, t_2)}{C^{(1)}[S_1(t_1), S_2(t_2)]\left\{1 - C^{(1)}[S_1(t_1), S_2(t_2)]\right\}}.$$

Therefore the asymptotic distribution of

$$\left(\frac{n}{m^{1/2}}\right)^{1/2} \left[\widehat{\mathrm{RR}}_m(t_1, t_2) - \mathrm{RR}(t_1, t_2)\right]$$

is the same as the asymptotic distribution of

$$\left(\frac{n}{m^{1/2}}\right)^{1/2} \frac{\mathrm{RR}(t_1, t_2)}{C^{(1)}[S_1(t_1), S_2(t_2)]\left\{1 - C^{(1)}[S_1(t_1), S_2(t_2)]\right\}} \times$$
$$\left\{C^{(1)}_{m,n}[S_{1n}(t_1), S_{2n}(t_2)] - C^{(1)}[S_1(t_1), S_2(t_2)]\right\},$$

which is, applying Lemma 2, $\mathcal{N}(0, \mathrm{Var}_{RR}(t_1, t_2))$.

5 Data applications

5.1 Example 1: Food expenditure and net income

In Abrams et al. (2020) we use a nonparametric Bernstein-based estimator for the cross ratio to explore the relationship between food expenditure and net income (Family Expenditure Survey, 1976, Härdle, 1990). We revisit this data example to illustrate the use of our novel nonparametric estimator for the risk ratio. As in Abrams et al. (2020) we use a random subsample of size n = 500 for our analysis. In Fig. D.1 of the Supplementary Material we graphically depict food expenditure (T_1) versus net income (T_2), where we express T_1 and T_2 in multiples of the expenditure sample mean, respectively, the net income sample mean (as suggested in Härdle, 1990). Table D.1 in the Supplementary Material provides the summary statistics.

In Fig. 6, we present the heatplot contourplot of the RR-surface. The risk ratio represents the ratio of the instantaneous probability of food expenditure equal to $T_1 = t_1$ given that the relative net income of individuals is larger than t_2 as compared to this probability for individuals with a relative net income smaller than t_2.

[Fig. 6 about here.]

In Fig. 7, we see that, when comparing households with a relative net income larger than t_2 to households with a relative net income smaller that t_2, the estimated risk ratio stays, over the entire food expenditure range, below one, at least when $t_2 = 0.978$ and $t_2 = 1.236$, and that in general the risk ratio is increasing as a function of food expenditure. We therefore have, given the direct link between the conditional hazards $\lambda(t_1|T_2 \geq t_2)$ and $\lambda(t_1|T_2 < t_2)$ and the corresponding conditional survival functions, that

$$P(T_1 > t_1|T_2 < t_2) < P(T_1 > t_1|T_2 \geq t_2).$$

This means that the higher income group typically spends more money for food when compared to the lower income group. Note that also for $t_2 = 0.545$ (the 25% empirical quantile for relative net income) this inequality remains valid in spite of the fact that the estimated risk ratio exceeds, for large values of t_1, the baseline value one ($\mathrm{RR}(t_1, t_2) \equiv 1$ means no association between T_1 and T_2). The reason being that for large t_1-values the baseline is only exceeded in a modest way, not strong enough to disrupt the inequality obtained when we conditioned on the 50% and 75% empirical quantile ($t_1 = 0.978$ and $t_2 = 1.236$). In general, the findings with regard to the relation between net income and food expenditure, as shown in Fig. 7, are in line with the association depicted in Figure 6 of the work by Abrams et al. (2020) and quantified in terms of the cross ratio function. Note, however, that the interpretation of the novel risk ratio is more straightforward than the CR, as it provides a direct comparison of the instantaneous risk of two groups in a dichotomised population (based on $\{T_2 \geq t_2\}$ and $\{T_2 < t_2\}$).

[Fig. 7 about here.]

Finally note that we used $m = 20$ as Bernstein order (see also the discussion in Section 6) given the moderate sample size and that we used an empirical version of the asymptotic variance in Theorem 1 to construct the 95% pointwise confidence bands in Figure 7.

5.2 Example 2: Cleveland heart disease data

In the second data example, we analyse the Cleveland heart disease dataset taken from the UCI Machine Learning repository (https://archive.ics.uci.edu/ ml/datasets/Heart+Disease). The data consist of patient information regarding 206 consecutive male patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984 (Detrano et al., 1989). The information included presence of heart disease, age, serum cholesterol (in mg/dl) and the maximum heart rate achieved by an individual (beats per minute). For more details concerning the data, the reader is referred to Appendix D of the Supplementary Material. Here, the relation between the maximum heart rate achieved and age is studied for males without ($n_1 = 92$) and with ($n_2 = 114$) heart disease. In Appendix D of the Supplementary Material we present results of the association between the serum cholesterol level and age, conditional on disease status.

In Fig. C.2 of the Supplementary Material, we graphically depict the age of the individual (T_2) versus the cholesterol level or the maximum heart rate achieved (T_1), respectively. The heatplots of the RR-surface for individuals with and without heart disease for the relation between age and maximum heart rate achieved are shown in Fig. C.5 of the Supplementary Material. The choice of the Bernstein order $m = 5$ in this analysis is inspired by the simulation results in case of a small sample size. In Figure 8, we show the estimated risk ratio as a function of the maximum heart rate achieved with age being larger than $t_2 = 47.0$ (25% percentile; upper panels), $t_2 = 54.5$ (50% percentile; middle panels) and $t_2 = 59.8$ (75% percentile; lower panels) with 95% asymptotic pointwise confidence bands. In patients without heart disease, the probability of a maximum heart rate larger than t_1 is larger for younger patients than for older ones. For patients suffering from heart disease, the estimated RR is close to one implying no association between age and maximum heart rate.

[Fig. 8 about here.]

6 Discussion

In this paper, we propose a Bernstein-based nonparametric estimator of the risk ratio $\mathrm{RR}(t_1, t_2)$, a new measure that unravels the local relation between the two components of (T_1, T_2). Simulations show, for interior points $(t_1, t_2) = (F_1^{-1}(u_1), F_2^{-1}(u_2))$ with (u_1, u_2) in the interior of $[0, 1] \times [0, 1]$, good finite sample performance of our estimator for different sample sizes. Also note that, since $S_{1n}(t_1) = 0$ and, in general $S_1(t_1) > 0$ for $t_1 > \max(t_{11}, t_{12}, \ldots, t_{1n})$, the estimator should not be used for t_1 values outside the data range. Moreover, it is well known that the asymptotic behaviour of the estimator in boundary points and in interior points is different, see, e.g., Janssen et al. (2016). We therefore focus in this paper on interior points, which are the points of practical value.

The use of Bernstein type estimators stems from the fact that

$$C_m(u, v) = \sum_{k=1}^{m} \sum_{l=1}^{m} C\left(\frac{k}{m}, \frac{l}{m}\right) P_{m,k}(u) P_{m,l}(v),$$

the Bernstein-Weierstrass approximation of $C(u, v)$, converges uniform to $C(u, v)$ as $m \to \infty$. Alternative copula-based type estimators for the risk ratio are possible. For example, estimators based on empirical beta copulas (Segers et al., 2017) or B-spline copulas (Shen et al., 2008) could be defined and studied. It would be nice to establish for these estimators asymptotic normality results that parallel our Theorem 1.

In many applications the variables T_1 and T_2 are survival times subject to right censoring. An interesting topic for future research is to define an estimator for the risk ratio $\mathrm{RR}(t_1, t_2)$ that uses Kaplan-Meier type estimators from survival analysis. A starting point could be the censored data estimators for the copula $C(\cdot, \cdot)$ as studied in Geerdens et al. (2016) or Gribkova and Lopez (2015). Note that extending the lemmas in Section 4 to the case of right-censored data is far from trivial and is beyond the scope of the present paper. A further step could be the incorporation of covariates into the estimation of the risk ratio. The methodological results in Section 4 provide appropriate guidelines to study these more complex data structures.

The inference we discussed for the risk ratio $\mathrm{RR}(t_1, t_2)$ can also be developed for risk ratios of form:

$$\frac{\lambda(t_1 \mid t_{21} \leq T_2 < t_{22})}{\lambda(t_1 \mid t_{23} \leq T_2 < t_{24})}.$$

This provides a lot of flexibility regarding the choice of the subgroups for which we want to study the risk ratio. Note that for $t_{21} = t_{24} = t_2, t_{23} = 0$ and $t_{22} \to \infty$, we obtain the risk ratio studied in this paper. Moreover, one can show that the cross ratio function CR defined in Section 1 is a limiting case. From this general form it is clear that the risk ratio is in general not symmetric in its arguments. Indeed, we rather think about T_2 as explanatory variable for T_1. Although the RR and CR are special cases of the aforementioned ratio of generalized conditional hazard functions, as pointed out earlier the RR offers a more straightforward interpretation than the CR. More specifically, since the conditioning on T_2 defines two mutually exclusive and comprehensive groups (i.e., defining a partition of the entire population) the risk ratio provides a direct comparison between the risks (or instantaneous probabilities) at time t_1 of both groups.

A further possible extension is to allow additional variables. In ongoing work the addition of a covariate X is considered, i.e.,

$$\frac{\lambda(t_1 | T_2 \geq t_2, X = x)}{\lambda(t_1 | T_2 < t_2, X = x)}.$$

This risk ratio can be expressed in terms of the conditional copula $C(.|X = x)$. Estimation of $C(.|X = x)$ has been studied in Gijbels et al. (2011) and Veraverbeke et al. (2011). In line with the present paper, these empirical copula estimators can be further smoothed using Bernstein polynomials. A further possible extension is obtained by replacing $X = x$ by $X \geq x$ (or $X < x$).

Also the case of bivariate 'primary' endpoints can be considered. Following, for example, Dabrowska (1988), we can study the conditional instantaneous rate of double failure at the point (t_1, t_2), i.e.,

$$\begin{aligned}\frac{\lambda(t_1, t_2 | T_3 \geq t_3)}{\lambda(t_1, t_2 | T_3 < t_3)} &= \frac{C^{(1,2)}\left[S_1(t_1), S_2(t_2), S_3(t_3)\right]}{C\left[S_1(t_1), S_2(t_2), S_3(t_3)\right]} \\ &\times \frac{C\left[S_1(t_1), S_2(t_2), 0\right] - C\left[S_1(t_1), S_2(t_2), S_3(t_3)\right]}{\left\{C^{(1,2)}\left[S_1(t_1), S_2(t_2), 0\right] - C^{(1,2)}\left[S_1(t_1), S_2(t_2), S_3(t_3)\right]\right\}}.\end{aligned}$$

Along the lines of the present paper, Bernstein type estimators can be proposed to estimate the copula and its derivatives to arrive at a nonparametric smoother to estimate the risk ratio.

Other variations on this risk ratio are within reach, i.e., the study of the risk ratio in case of bivariate endpoints that looks at the conditional instantaneous risk of single failure at t_1 (or at t_2) (see Dabrowska, 1988, for details on the definition).

Finally, it would be nice to determine the optimal Bernstein order m in a data-driven way. Based on our simulations, increasing the Bernstein order with increasing sample size seems a logical choice. We expect that the optimal order also depends on the underlying association structure. Bootstrap-based and/or cross-validation methods could be useful to select, for a specific dataset, the appropriate Bernstein order m. As far as we know, only a few papers appeared so far on data-driven choices of the order m. Taylor-Rodriguez and Ghosh (2021) look at the problem in a regression context and in the recent master thesis of Sachithra-Opathalage (2021) selection of the order is discussed for density estimation.

Disclosure statement

No potential competing interest was reported by the authors.

Data availability statement

The heart disease data that support the findings of this study are available in the UCI Machine Learning repository at https://archive.ics.uci.edu/ ml/datasets/Heart+Disease. The generated food expenditure data is publically available at www.ibiostat.be.

References

Abrams S, Janssen P, Swanepoel J, Veraverbeke N (2020) Nonparametric estimation of the cross ratio function. Annals of the Institute of Statistical Mathematics 72:771–801

Arnold BC, Kim YH (1996) Conditional proportional hazards models. Lifetime Data: Models in Reliability and Survival Analysis pp 21–28

Bouezmarni T, Rombouts K, Taamouti A (2009) Asymptotic properties of the Bernstein density copula estimator for α-mixing data. Journal of Multivariate Analysis 101:1–10

Bouezmarni T, El Ghouch A, Taamouti A (2013) Bernstein estimator for unbounded copula densities. Statistics and Risk Modeling 30:343–360

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65:141–151

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, Boca Raton

Dabrowska DM (1988) Kaplan-Meier estimate on the plane. Annals of Statistics 16(4):1475–1489

Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid J, Sandhu S, Guppy K, Lee S, Froelicher V (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology 64:304–310

Fermanian JD, Radulovic D, Wegkamp M (2004) Weak convergence of empirical copula processes. Bernoulli 10:847–860

Geerdens C, Janssen P, Veraverbeke N (2016) Large sample properties of nonparametric copula estimators under bivariate censoring. Statistics 50:1036–1055

Gijbels I, Veraverbeke N, Omelka M (2011) Conditional copulas, association measures and their applications. Computational Statistics & Data Analysis 55(5):1919–1932

Gribkova S, Lopez O (2015) Non-parametric copula estimation under bivariate censoring. Scandinavian Journal of Statistics 42(4):925–946

Härdle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge

Janssen P, Swanepoel J, Veraverbeke N (2012) Large sample behavior of the Bernstein copula estimator. Journal of Statistical Planning and Inference 142:1189–1197

Janssen P, Swanepoel J, Veraverbeke N (2014) A note on the asymptotic behavior of the Bernstein estimator of the copula density. Journal of Multivariate Analysis 124:480–487

Janssen P, Swanepoel J, Veraverbeke N (2016) Bernstein estimation for a copula derivative with application to conditional distribution and regression functionals. Test 25:351–374

Leblanc A (2012) On estimating distribution functions using Bernstein polynomials. Annals of the Institute of Statistical Mathematics 64:919–943

Nelsen RB (2006) An introduction to copulas. Springer, New York

Sachithra-Opathalage G (2021) Data-driven smoothing parameter selection in density estimation. Master's thesis, University of Manitoba

Sancetta A, Satchell S (2004) The Bernstein copula and its applications to modeling and approximations of multivariate distributions. Economic Theory 20:535–562

Segers J, Sibuya M, Tsukahara H (2017) The empirical beta copula. Journal of Multivariate Analysis 155:35–51

Shen X, Zhu Y, Song L (2008) Linear B-spline copulas with application to nonparametric estimation of copulas. Computational Statistics and Data Analysis 52(7):3806–3819

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publications de l'institut de statistique de l'Université de Paris 8:229–231

Taylor-Rodriguez D, Ghosh S (2021) On the estimation of the order of smoothness of the regression function. Tech. rep.

Veraverbeke N, Omelka M, Gijbels I (2011) Estimation of a conditional copula and association measures. Scandinavian Journal of Statistics 38(4):766–780

List of Figures

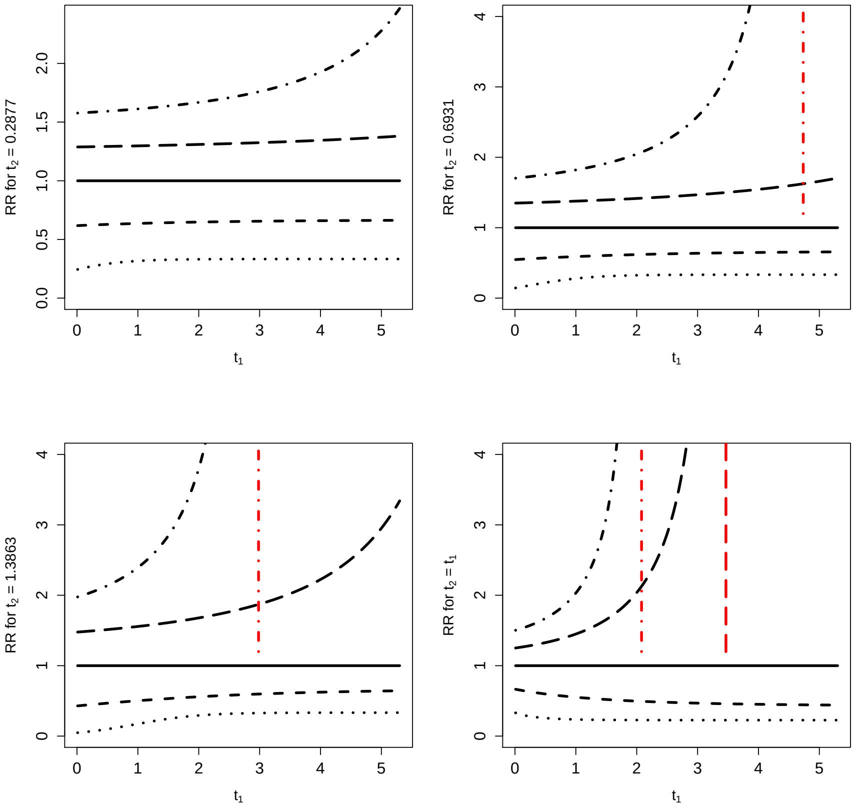

Fig. 1 Risk ratio curves for a Clayton copula function with exponential event times T_1 and T_2 ($\lambda_1 = \lambda_2 = 1$), for $\theta = -1/3$ (dash-dotted lines), $\theta = -0.2$ (large dashed lines), $\theta = 0$ (solid lines), $\theta = 0.5$ (dashed lines), and $\theta = 2$ (dotted lines), and for $t_2 = 0.2877$ (left upper panel), $t_2 = 0.6931$ (right upper panel), $t_2 = 1.3863$ (left lower panel) and $t_2 = t_1$ (right lower panel). Vertical red lines indicate the boundary constraints on the domain of the risk ratio function $\mathrm{RR}(t_1, t_2)$ for negative parameter values (i.e., $\theta = -1/3$ (red dash-dotted line) and $\theta = -0.2$ (red large dashed lines)).

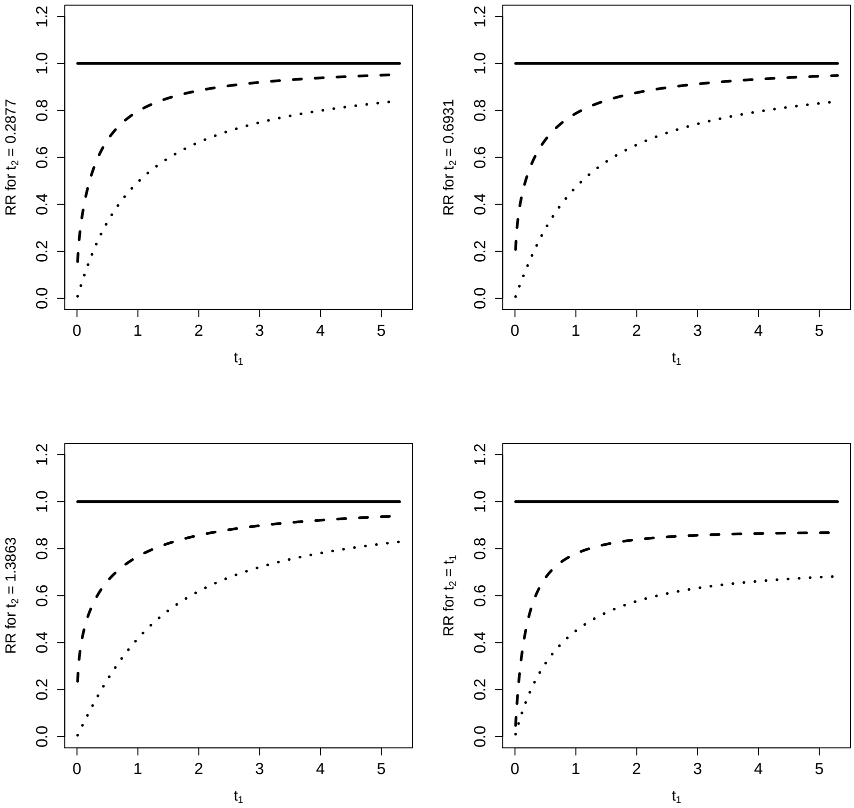

Fig. 2 Risk ratio curves for a Gumbel copula function with exponential event times T_1 and T_2 ($\lambda_1 = \lambda_2 = 1$), for $\theta = 1$ (solid lines), $\theta = 10/8$ (dashed lines), and $\theta = 2$ (dotted lines), and for $t_2 = 0.2877$ (left upper panel), $t_2 = 0.6931$ (right upper panel), $t_2 = 1.3863$ (left lower panel) and $t_2 = t_1$ (right lower panel).

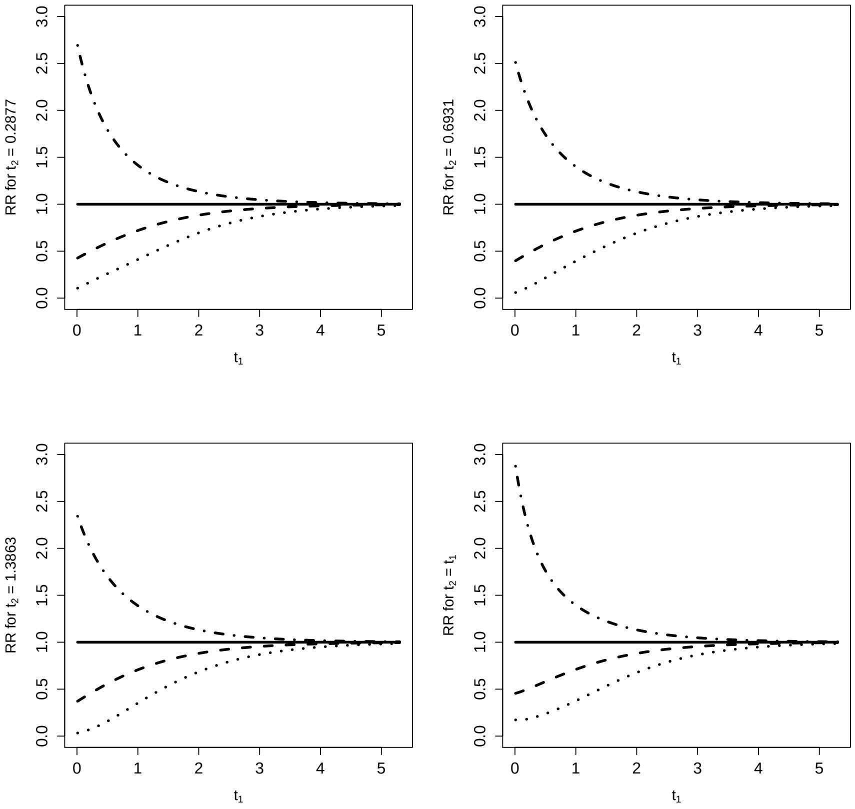

Fig. 3 Risk ratio curves for a Frank copula function with exponential event times T_1 and T_2 ($\lambda_1 = \lambda_2 = 1$), for $\theta = -1.8609$ (dash-dotted lines), $\theta = 10^{-6}$ (solid lines), $\theta = 1.8609$ (dashed lines), and $\theta = 5.7363$ (dotted lines), and for $t_2 = 0.2877$ (left upper panel), $t_2 = 0.6931$ (right upper panel), $t_2 = 1.3863$ (left lower panel) and $t_2 = t_1$ (right lower panel).

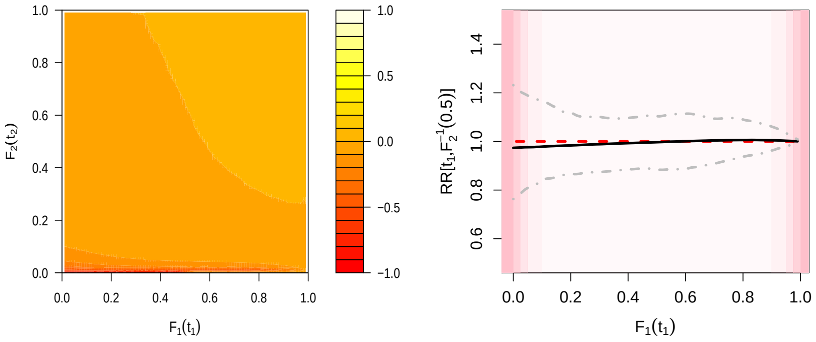

Fig. 4 Clayton copula with $\theta = 0$ (independence model), $m = 5$ and $n = 500$: heatplot representing the relative difference between the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ averaged over the M replications and the true risk ratio $\mathrm{RR}(t_1, t_2)$ (left panel) and the estimated risk ratio as a function of t_1 with $t_2 = F_2^{-1}(0.5)$ fixed (right panel; black solid line) and with pointwise 95% simulation-based confidence bounds (gray dash-dotted lines). The red dashed line represents the true risk ratio.

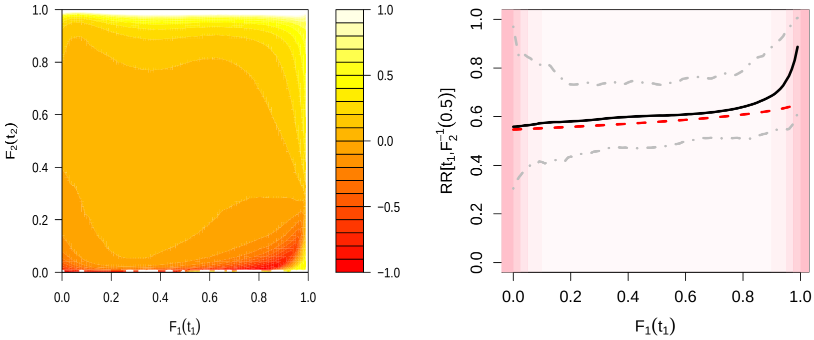

Fig. 5 Clayton copula with $\theta = 0.5$, $m = 20$ and $n = 500$: heatplot representing the relative difference between the estimated risk ratio $\widehat{\mathrm{RR}}_m(t_1, t_2)$ averaged over the M replications and the true risk ratio $\mathrm{RR}(t_1, t_2)$ (left panel) and the estimated risk ratio as a function of t_1 with $t_2 = F_2^{-1}(0.5)$ fixed (right panel; black solid line) and with pointwise 95% simulation-based confidence bounds (gray dash-dotted lines). The red dashed line represents the true risk ratio.

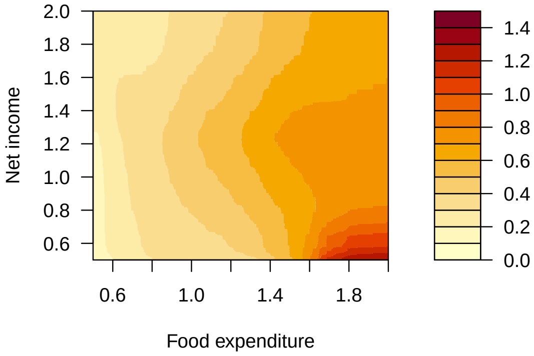

Fig. 6 Food expenditure and net income data application: heatplot of the estimated food expenditure risk ratio ($m = 20$).

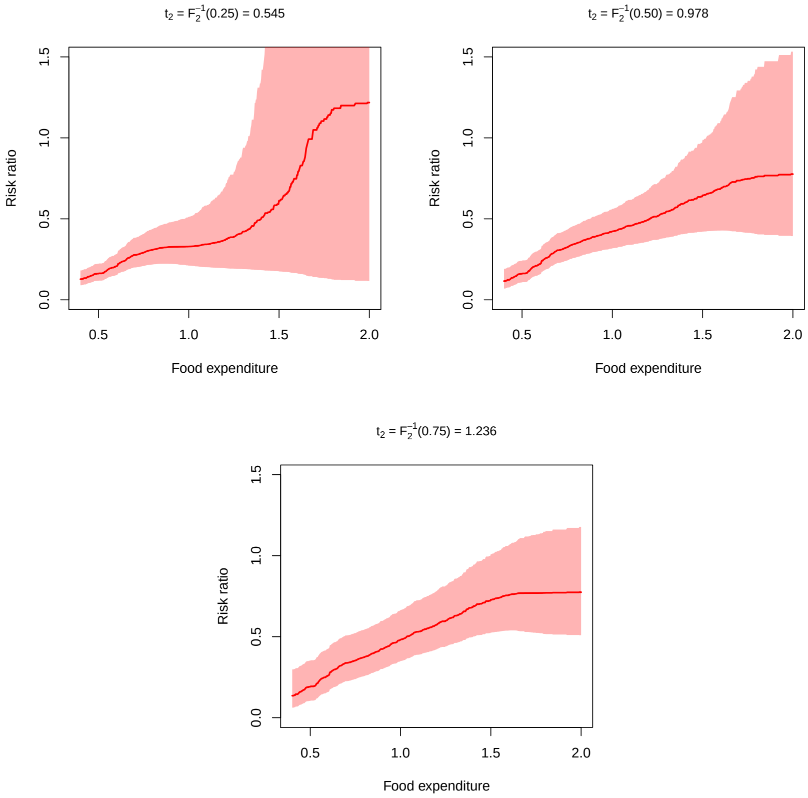

Fig. 7 The estimated risk ratio as a function of t_1 ($m = 20$) for relative net income $t_2 = 0.545$ (left upper panel), $t_2 = 0.978$ (right upper panel) or $t_2 = 1.236$ (lower panel) fixed and with pointwise asymptotic 95% confidence bounds (shaded areas).

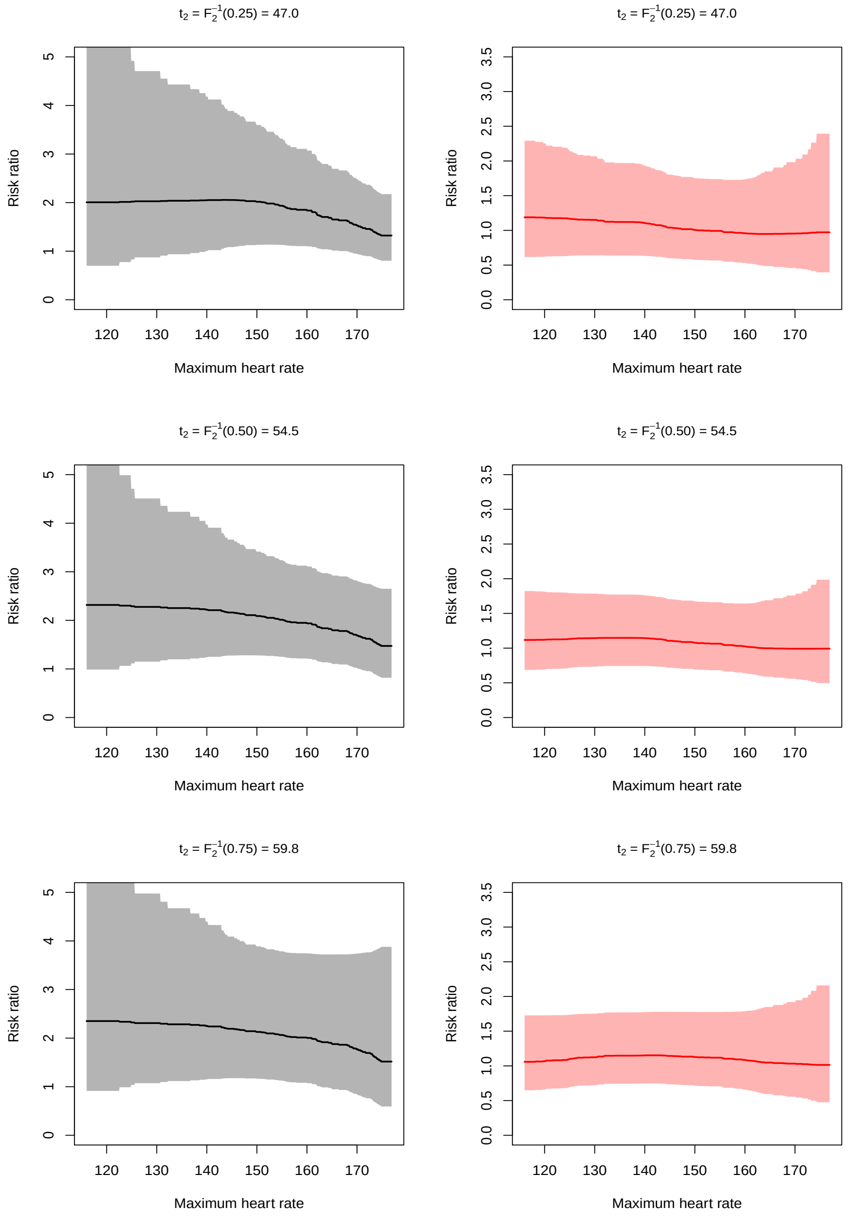

Fig. 8 The estimated risk ratio as a function of the maximum heart rate achieved t_1 ($m = 5$) for age $t_2 = 47.0$ (upper panels), $t_2 = 54.5$ (middle panels) or $t_2 = 59.8$ years (lower panels) fixed in patients without (left panels; black lines) and with heart disease (right panels; red lines) and with pointwise asymptotic 95% confidence bounds (shaded areas).

List of Tables

Table 1 The upper boundaries for t_1 for different values of $\theta < 0$ in a Clayton copula setting.

t_2 \ θ	$-1/3$	-0.2
0.2877	7.1861	14.4195
0.6931	4.7429	10.2226
1.3863	2.9878	7.0911

Table 2 MI_{RR}, ISB_{RR} and IV_{RR} for different choices of m and different assumptions regarding the dependence between the event times; Clayton copula function with parameter values for θ equal to 0.00 (independence), 0.50 and 2.00. Minimum MI_{RR} values are indicated in bold.

		Independence ($\theta = 0.00$, $\tau = 0.00$)									
n	Measure	5	10	15	20	25	30	35	40	45	50
300	MI_{RR}	**0.033**	0.083	0.133	0.179	0.224	0.282	0.353	0.351	0.347	0.443
	ISB_{RR}	0.003	0.003	0.005	0.005	0.004	0.005	0.004	0.010	0.006	0.007
	IV_{RR}	0.030	0.079	0.128	0.174	0.219	0.277	0.348	0.341	0.341	0.436
500	MI_{RR}	**0.018**	0.049	0.089	0.123	0.144	0.179	0.181	0.203	0.234	0.274
	ISB_{RR}	0.001	0.001	0.002	0.001	0.001	0.006	0.003	0.001	0.003	0.006
	IV_{RR}	0.018	0.049	0.087	0.122	0.142	0.173	0.178	0.202	0.231	0.267
800	MI_{RR}	**0.011**	0.030	0.054	0.067	0.076	0.111	0.109	0.126	0.134	0.148
	ISB_{RR}	0.000	0.001	0.000	0.001	0.001	0.001	0.001	0.002	0.001	0.002
	IV_{RR}	0.010	0.029	0.053	0.066	0.075	0.111	0.108	0.125	0.133	0.147

		Clayton copula ($\theta = 0.50$, $\tau = 0.20$)									
n	Measure	5	10	15	20	25	30	35	40	45	50
300	MI_{RR}	0.323	0.114	0.092	**0.092**	0.099	0.098	0.116	0.117	0.127	0.131
	ISB_{RR}	0.307	0.079	0.045	0.025	0.020	0.016	0.011	0.010	0.008	0.012
	IV_{RR}	0.016	0.035	0.048	0.067	0.078	0.081	0.105	0.107	0.119	0.119
500	MI_{RR}	0.321	0.110	0.067	**0.054**	0.057	0.055	0.061	0.067	0.072	0.079
	ISB_{RR}	0.312	0.087	0.038	0.020	0.016	0.010	0.006	0.006	0.005	0.005
	IV_{RR}	0.009	0.022	0.029	0.035	0.041	0.045	0.054	0.061	0.067	0.074
800	MI_{RR}	0.305	0.110	0.059	0.046	0.044	**0.036**	0.047	0.050	0.049	0.050
	ISB_{RR}	0.299	0.096	0.040	0.022	0.015	0.008	0.009	0.007	0.005	0.003
	IV_{RR}	0.006	0.015	0.019	0.024	0.029	0.028	0.038	0.044	0.044	0.047

		Clayton copula ($\theta = 2.00$, $\tau = 0.50$)									
n	Measure	5	10	15	20	25	30	35	40	45	50
300	MI_{RR}	0.606	0.194	0.114	0.090	0.083	0.081	**0.071**	0.075	0.080	0.084
	ISB_{RR}	0.596	0.174	0.086	0.059	0.051	0.047	0.030	0.037	0.031	0.046
	IV_{RR}	0.011	0.021	0.028	0.030	0.033	0.034	0.042	0.038	0.049	0.037
500	MI_{RR}	0.611	0.187	0.094	0.068	0.058	0.049	0.049	**0.048**	0.048	0.053
	ISB_{RR}	0.606	0.177	0.083	0.055	0.043	0.031	0.023	0.028	0.020	0.034
	IV_{RR}	0.005	0.009	0.011	0.013	0.015	0.018	0.027	0.020	0.028	0.020
800	MI_{RR}	0.622	0.186	0.090	0.058	0.045	0.037	0.039	0.037	**0.031**	0.036
	ISB_{RR}	0.619	0.180	0.083	0.050	0.036	0.025	0.018	0.025	0.015	0.023
	IV_{RR}	0.004	0.006	0.007	0.008	0.009	0.012	0.021	0.012	0.016	0.013