Received: 1 March 2019 | Revised: 20 September 2019 | Accepted: 21 September 2019

DOI: 10.1002/sim.8393

RESEARCH ARTICLE

WILEY Statistics in Medicine

Multistate quantile regression models

Alessio Farcomeni[1] | **Marco Geraci**[2]

[1]Department of Economics and Finance, University of Rome "Tor Vergata", Rome, Italy

[2]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina

Correspondence
Alessio Farcomeni, Department of Economics and Finance, University of Rome "Tor Vergata", Via Columbia 2, 00133 Roma, Italy.
Email: alessio.farcomeni@uniroma2.it

We develop regression methods for inference on conditional quantiles of time-to-transition in multistate processes. Special cases include survival, recurrent event, semicompeting, and competing risk data. We use an ad hoc representation of the underlying stochastic process, in conjunction with methods for censored quantile regression. In a simulation study, we demonstrate that the proposed approach has a superior finite sample performance over simple methods for censored quantile regression, which naively assume independence between states, and over methods for competing risks, even when the latter are applied to competing risk data settings. We apply our approach to data on hospital-acquired infections in cirrhotic patients, showing a quantile-dependent effect of catheterization on time to infection.

KEYWORDS

censored quantiles, cross-infection, duration models

1 | INTRODUCTION

Traditional survival models (eg, Cox regression) are able to deal with a single or composite endpoint. However, in some studies, several endpoints are present. For instance, in a breast cancer trial, possible outcomes include disease-free survival, local recurrence, distant metastasis, or death. In such cases, separate analyses are sometimes carried out for each of the endpoints. These separate analyses can be biased[1] and they fail to reveal the relations between different types of events.[2,3] In recent years, multistate models (MSMs)[1] have gained a prominent role in clinical and epidemiological studies.[4-8] MSMs are an extension of the classical survival model for analyzing complex time-to-event problems with multiple endpoints. Compared to models with a single or a composite endpoint, MSMs have several advantages. First of all, they remove potential sources of bias.[1] Secondly, etiological aspects of different phases of the disease can be studied. Analysis of competing states is possible, such as different causes of death or competing therapy outcomes or a sequence of states such as disease recurrences. Finally, from a predictive point of view, prognosis from MSMs can be more accurate than from the standard model with one single endpoint.[1,9] However, existing MSMs are typically based on strong assumptions and the classical hazard-based interpretation of the results may be difficult.

Quantile regression (QR)[10] is a nonparametric alternative to classical location-shift models, which aims at modeling the impact of covariates on quantiles of a response variable. QR has become a successful analytic method in many fields of science because of its ability to draw inferences about individuals that rank below or above the population conditional mean. The ranking within the conditional distribution of the outcome can be considered as a natural index of individual latent characteristics, which cause heterogeneity at the population level.[11] This is particularly relevant in survival analysis since treatments may have different (even opposite) effects on different quantiles. Moreover, the conditional median is an excellent alternative to the conditional mean when the error is skewed. This advantage is of particular relevance if we consider that the mean (and other moments) may not even be identified with censored data, while quantiles of some order can always be estimated (as long as not all observations are censored).

The literature on QR methods for survival data is quite varied. There are proposals to estimate censored QR parameters using a modified Kaplan-Meier estimator,[12] martingale-based estimating equations,[13] as well as weighted estimating

equations.[14-17] Recently, data augmentation[18] and a modification of the check loss function[19] for censoring have been proposed. More in general, time-to-event data comprise possibly complex settings, such as recurrent events, competing risks, and semicompeting risks. Individuals may experience the same, nonterminal event multiple times (recurrent events). Competing risks arise when different types of events are possible, and they are all terminal, while semicompeting risks have two event: one terminal and one nonterminal. QR approaches have been developed for such scenarios.[20-26]

In this work, we are interested in QR for discrete state-space continuous-time stochastic processes, a class of multistate processes.[1,27] In our setting, there are a number of states (eg, healthy, transplanted, graft failed, and dead) and individuals jump from a given state to another. Jumps occur at random times, with probabilities that depend on a set of covariates. When an individual jumps to a particular state, unrealized transitions to alternative states are considered to be censored. In our analytic approach, we allow for the following: (i) the exclusion of particular observation times if these are related to transitions that are not of interest; (ii) one or more absorbing states, that is, terminal states which, once reached, cannot be departed from (eg, death); (iii) states to which transitions are impossible (eg, from healthy to graft failure); and (iv) repeated measures on the same subjects who visit the same state multiple times. With r considered both conditional and marginal models to deal with repeated measures in QR.[28] In the former category, we find random effects models, either linear[29-33] or nonlinear,[34,35] which, however, can be computationally demanding. In the marginal models category, the weighted approach[36] is perhaps the easiest to implement. Ignoring dependency is nonetheless a successful strategy.[24] This strategy is linked to marginal modeling[37] and leads to a consistent estimator under quite general assumptions.[38,39] Copula models have also been used in QR modeling with time-to-event data.[23,25]

In summary, we develop an approach to QR to model time-to-transition following a first-jump representation of the data, via censored QR. Our methods are general, as they can be applied to different multistate processes including, but not limited to, recurrent events, competing risks, and semicompeting risks, and they can be implemented using readily available software. In a simulation study, we show that our estimator has excellent properties in terms of mean squared error for both predictions and parameters. Notably, our estimator outperforms existing QR methods for competing risk, even when the latter is applied to the analysis of competing risk data. We also formally discuss assessment of homogeneity assumptions ("state merging"). This procedure implies that all the transitions from (or to) states within the same group are assumed to result from the same process, with equal probability. In order to test this assumption, we derive a Wald-type test statistic. Our proposal is motivated by an application to risk modeling for hospital-acquired infections and related deaths in Italian cirrhotic patients. Using a retrospective database of 870 patients, some of whom experienced multiple hospitalizations, we estimate quantiles of time-to-infection and time-to-death associated with catheterization and paracentesis, after adjusting for potential confounders. Our results indicate that catheterization reduces the time-to-infection (ie, increases the risk of early infection) and that the magnitude of the effect increases with quantile level, while there is weak evidence supporting the negative effect of paracentesis.

The rest of this paper is organized as follows. In the next section, we present our methodology and discuss related issues, including homogeneity constraints and testing. In Section 3, we report the results of a simulation study and, in Section 4, the analysis of the data example. Some concluding remarks are given in Section 5.

2 | METHODS

2.1 | Model and estimation

Let X_t, $t \geq 0$, be a discrete state-space continuous-time stochastic process with state-space $\{1, \ldots, s\}$ and $s \geq 2$. The classical framework of MSMs is based on time-dependent transition probabilities and cannot be linked directly to QR given that these probabilities do not form a cumulative distribution function (CDF). However, one can represent the stochastic process X_t according to first hitting times. Suppose state m is not an absorbing state and let $T_m = \inf\{t \geq 0 : X_t \neq m, X_0 = m\}$ be the random variable that defines the time at first jump from state m. Also, let U_j denote the sequence of states that the continuous time Markov chain visits (regardless of dwelling times), with U_0 being the initial state. Then, due to the Markov homogeneity assumptions, the stochastic process is completely specified by

$$F_{mj}(t) = \Pr(T_m \leq t, U_1 = j \mid U_0 = m) = \Pr(T_{mj} \leq t), \quad \text{for } t > 0, \tag{1}$$

where T_{mj} is the time to transition from state m to state j, $m = 1, \ldots, s$, and $j \neq m$.

From state	To state	Initial time	Final time	$\widetilde{T}_{12}$	Δ_{12}
1	2	0	3.4	3.4	1
2	1	3.4	5	NA	NA
1	3	5	5.8	∞	0
3	1	5.8	6.2	NA	NA
1	2	6.2	8	1.8	1
2	1	8	8.5	NA	NA
1	0	8.5	10	1.5	0

TABLE 1 Hypothetical transition history for a subject, where the transition of interest is from state 1 to state 2. State 0 indicates censoring

For practical purposes, we do not work directly with T_{mj}. Instead, we define the variable $T^*_{mj} = T_{mj}$ if the transition is to state j and $T^*_{mj} = \infty$ if transition is to some other state. When the follow-up is closed in state m, then we have an (independent) censoring time C_m, with $C_m = \infty$ if a transition occurs. Hence, the variables relevant to the analysis are $\widetilde{T}_{mj} = \min(T^*_{mj}, C_m)$ and $\Delta_{mj} = I(T^*_{mj} < C_m)$. Moreover, let $Z = (Z_1, Z_2, \ldots, Z_p)'$ denote a $p \times 1$ vector of predictors. We can now introduce the conditional (on Z) quantile function

$$Q_{mj}(\tau \mid Z) = \inf\{t : \Pr(T^*_{mj} \leq t, U_1 = j \mid Z, U_0 = m) \geq \tau\},$$

$\tau \in (0, 1)$, which is the left inverse of the conditional CDF $\Pr(T^*_{mj} \leq t, U_1 = j \mid Z, U_0 = m)$. To make explicit the relationship between $Q_{mj}(\tau \mid Z)$ and Z for a given τ, we specify a transformation model of the kind

$$Q_{mj}(\tau \mid Z) = g_{mj}\{\alpha(\tau) + Z'\beta_{mj}(\tau)\}, \tag{2}$$

where $g_{mj}(\cdot)$ is a known monotone link function, $\alpha(\tau)$ is a quantile-specific intercept, and $\beta_{mj}(\tau) = (\beta_{mj1}(\tau), \beta_{mj2}(\tau), \ldots, \beta_{mjp}(\tau))'$ is a $p \times 1$ vector of quantile-specific regression coefficients. The kth element of $\beta_{mj}(\tau)$ is denoted with $\beta_{mjk}(\tau)$, $k = 1, \ldots, p$, and can be interpreted as the change in the τth quantile of T^*_{mj}, on the scale of $g_{mj}(\cdot)$, when Z_k is incremented by one unit and all other predictors are held fixed. To simplify interpretation of parameters, the same link function will be specified for any m and j, although in principle, one can use different transformations.

In summary, suppose the transition of interest is from state m to state j. Transitions from the state of interest m to any other state $h \neq j$ are replaced with an infinite duration time and censored. Transitions from any state $h \neq m$ are removed from the current estimation set. On the other hand, repeated transitions from m to j enter the estimation set as multiple events. To fix the ideas, consider a subject with observation history as in Table 1, and suppose transitions from state 1 to state 2 are of interest. It is clear that, even if the subject has seven transitions, three of those are excluded from the estimation set and only two of those qualify as an event.

The conditional quantiles $Q_{mj}(\tau \mid Z)$, as well as the marginal quantiles $Q_{mj}(\tau)$, can be estimated by means of any method for censored quantile regression, which sets the framework of the ensuing inferential approach. In the remainder of this paper, we rely on Portnoy's[12] estimator, which, in turn, is based on a direct extension of Kaplan-Meier–type estimators. Similarly, there exist a variety of approaches to deal with repeated measurements as briefly summarized in our introductory remarks. In Sections 3 and 4, we relied on the consistency of the estimator,[38,39] which essentially correspond to Karlsson's[36] approach with uniform weights. This is linked to a marginal modeling approach for repeated measures. In our simulation study, we did try several other weight specifications, all of which performed somewhat poorly in terms of mean squared error (results not shown). We speculate that weight estimation entails loss of efficiency. In the remainder of this paper, by "multistate quantile regression" (MSQR) estimator, we mean the estimator of the parameters in (2) that relies on the combination of first-hitting times representation of the data, uniform weights, and Portnoy's[12] estimation.

Finally, the implementation of the proposed methods necessitates defining an interval of estimable τ's as, in practice, the data provide limited information to the QR estimator. To our knowledge, this problem has not yet been formally tackled in the literature of quantiles for censored data. While it is beyond the scope of this paper to develop a general approach, we propose the following ad hoc strategy. Empirical cumulative probabilities are estimated for different covariate configurations (this requires discretizing continuous covariates as appropriate). An upper bound for the range of estimable quantiles can be defined as the smallest upper bound of the stratified cumulative probability curves. On the other hand, the lower bound can be selected as the minimal quantile for which all stratified cumulative probability curves are not clearly following a model different than (2). A similar approach has already been explored for competing risk QR.[21]

2.2 | Homogeneity constraints and testing

In multistate modeling, merging two or more states into a single one is a common practice. This is done either because the differences between specific states are practically irrelevant or considered to be of little scientific interest, or to obviate lack of a sufficient number of events for some transitions. As a result, covariate effects are assumed to be homogeneous across certain transitions.

Homogeneity constraints are usually tested through likelihood ratio statistics.[40] However, in our setting, we prefer Wald-type tests based on parameter estimates, with the inverse Fisher information replaced by an estimate of the covariance matrix. We chose this approach since several nonparametric methods for censored quantile regression are not even associated with a likelihood, and even working likelihoods do not usually arise from the data-generating mechanism. Another reason for our preference is that, to be computed, our test statistics only require fitting the unconstrained model. The constrained model is fitted if the test does not lead to rejection of the null hypothesis of homogeneity.

We want to test the null hypothesis $H_0 : \beta_{m_1 j_1 k_1}(\tau) = \cdots = \beta_{m_r j_r k_r}(\tau)$ for some $r \geq 2$, where $\beta_{m_h j_h k_h}(\tau)$ is the regression coefficient associated with the predictor Z_{k_h} for transition from state m_h to j_h. Note that we allow for testing homogeneity of coefficients associated both with the same ($k_1 = k_2 = \cdots = k_r$) or with different (at least one k_h is different than the other ones) predictors. We define the test statistic

$$W_h = \frac{\{\widehat{\beta}_{m_h j_h k_h}(\tau) - \widehat{\beta}_{m_r j_r k_r}(\tau)\}^2}{s_{hh} + s_{rr} + 2s_{hr}}, \quad h = 1, \ldots, r-1, \tag{3}$$

where s_{ab} denotes the abth element of the $r \times r$ variance-covariance matrix $S = \text{cov}\{\widehat{\beta}_{m_1 j_1 k_1}(\tau), \ldots, \widehat{\beta}_{m_r j_r k_r}(\tau)\}$. Note that the particular choice of $\widehat{\beta}_{m_r j_r k_r}$ to obtain the $r-1$ contrasts is arbitrary. We consider the statistic

$$W = \sum_{h=1}^{r-1} W_h,$$

which, under the null hypothesis, is approximately χ^2_{r-1}. Finally, we propose estimating S by block-bootstrap. More specifically, a sample of size n is drawn with replacement from the current data. When the ith subject is sampled, their entire transition history is included in the new sample. Sampling and estimation of parameters for all transitions and quantiles of interest is repeated a large number of times and the parameter's covariance estimate is obtained as the sample covariance of the bootstrap estimates.

Partial homogeneity constraints are also easily tackled in our framework. By partial homogeneity, we mean the situation in which only the effect of certain covariates is homogeneous across multiple transition types, while other effects are heterogeneous. In order to estimate a model under partial homogeneity constraints, we pool observations relating to all transitions of interest, and treat them as if they were the same transition. Successively, we include interactions between unconstrained effects and transition-type indicators in the regression model.

3 | SIMULATION STUDY

In this section, we investigate the finite-sample properties of our proposed multistate quantile regression (MSQR) estimator. The data were generated ($B = 1000$ replications) under two scenarios: a competing risk scenario and a more general multistate scenario. Each scenario was considered with either three or five states and the error was generated from either a log-normal or a Weibull distribution. A summary of the settings used is given in Table 2. A detailed description of the data generating mechanisms in each scenario is given in the following.

TABLE 2 Summary of the settings used for the competing risks and multistate scenarios. For each scenario, there are 2 cardinalities of state sets, 2 error distributions, and 4 combinations of γ and c_u, for a total of $2 \times 2 \times 4 = 16$ distinct cases

	Competing risks				Multistate			
Number of states	**3**		**5**		**3**		**5**	
Absorbing states	{2, 3}		{2, 3, 4, 5}		{3}			
Error	Log-normal or Weibull				Log-normal or Weibull			
γ	−1	0	−1	0	−1	0	−1	0
c_u	5	5	8	8	5	5	8	8

3.1 | Competing risk scenario

Competing risk data with three states were generated similarly to those by Peng and Fine,[21] with an initial state 1 and absorbing states 2 and 3. Subjects move from state 1 with probability depending on two covariates Z_1 and Z_2, generated as a standard uniform and a Bernoulli with probability 50%, respectively. The distribution of the censoring indicator is a mixture with a point mass at c_u, where c_u is either 5 or 8. Therefore, censoring has conditional probability $0.8t/c_u$ for $t < c_u$ and 1 for $t \geq c_u$. Transition indicators are generated according to $\Pr(U_t = 2 \mid Z, U_{t-1} = 1) = 0.8I(Z_2 = 0) + 0.6I(Z_2 = 1)$.

In one specification of the model's error, the time-to-event distribution was defined as

$$\Pr(T \leq t \mid U_t = j, U_{t-1} = i, Z) = \Phi(\log t - \gamma_{ij}' Z),$$

where $\gamma_{12} = (\gamma, -0.5)$ and $\gamma_{13} = (0, -0.5)$. Data were simulated with γ set at either -1 or 0. In another specification of the model's error, the time-to-event distribution, conditionally on $U_t = j, U_{t-1} = i$, and Z, was defined according to a Weibull distribution with scale and shape parameters equal to, respectively, $\frac{1}{4} + |\gamma_{ij}' Z|$ and 0.5 (while all the other parameters were fixed at the same values as above). Note that the link function is exponential in the first case, while it is the identity under the Weibull.

When simulating a process with five states, all states except state 1 were absorbing. The other parameters were set as follows: $\Pr(U_t = j \mid U_{t-1} = 1, Z_2 = 0) = 0.25$ for $j = 2, 3, 4, 5$, $\Pr(U_t = 2 \mid U_{t-1} = 1, Z_2 = 1) = 0.7$, $\Pr(U_t = j \mid U_{t-1} = 1, Z_2 = 1) = 0.1$ for $j = 3, 4, 5$, $\gamma_{14} = (1, -0.5)$, and $\gamma_{15} = (-1, -0.5)$.

3.2 | Multistate process scenario

In the multistate scenario, data were generated as in the competing risk scenario, except that now, only state 3 is absorbing and

$$\Pr(U_t = 1 \mid Z, U_{t-1} = 2) = 0.5I(Z_2 = 0) + 0.4I(Z_2 = 1).$$

Similarly to the competing risk scenario, the same time-to-event distributions (log-normal or Weibull) were used, with, additionally, $\gamma_{21} = (0.25, 0.75)$ and $\gamma_{23} = (2, -1)$.

When simulating a process with 5 states, the following parameters were used: $\Pr(U_t = 1 \mid U_{t-1} = 2, Z_2 = 0) = 0.25$, $\Pr(U_t = 1 \mid U_{t-1} = 2, Z_2 = 1) = 0.1$, $\gamma_{21} = (0.25, 0.75)$, $\gamma_{23} = (2, -1)$, $\gamma_{24} = (-1, 2)$, and $\gamma_{25} = (0, 0.5)$.

3.3 | Performance criteria and results

We estimated parameters and predicted conditional quantiles at levels $\tau = 0.1$ and $\tau = 0.5$ for the transition from state 1 to state 2 using three methods: our proposed MSQR estimator; a naïve approach that assumes independence between states based on standard censored quantile regression (CRQ),[12] which is implemented in the function `crq` from the `R` package `quantreg`; and one approach for competing risk quantile regression cmprskQR,[21] which is implemented in the function `crrQR` from the `R` package `cmprskQR`. When applying the CRQ approach, we treated competing events as censoring. In contrast, when applying the method of Peng and Fine in noncompeting risk settings, we (inappropriately) treated transient events as censoring.

After fitting models, we calculated the root mean squared error (RMSE) and bias for both $\widehat{\beta}(\tau)$ and $\widehat{Q}(\tau)$. For the former, RMSE and bias were calculated as the average of the elementwise RMSE and bias estimates, respectively. Under log-normal errors, one can verify that the parameter to be estimated is

$$\beta_{12}(\tau) = \left[\Phi^{-1}\left(\frac{\tau}{0.8}\right), \gamma, -0.5 + \Phi^{-1}\left(\frac{\tau}{0.6}\right) - \Phi^{-1}\left(\frac{\tau}{0.8}\right)\right]'.$$

Under Weibull errors, the true value of $\beta_{12}(\tau)$ was numerically approximated by first generating 10^4 observations with no censoring and then fitting a linear quantile model.[10]

To save space, here, we show only selected results for $\widehat{\beta}(\tau)$ when times are log-normal. In Figures 1 and 2, we report boxplots of RMSE and bias of each estimator under the competing risk scenario, and, in Figures 3 and 4, under the multistate process scenario. All the other results are shown in supplementary materials.

The MSQR estimator outperforms, often considerably, both the CRQ and the methods of Peng and Fine almost at every combination of sample size and parameters. Remarkably, our estimator is superior to the competing risk estimator of Peng

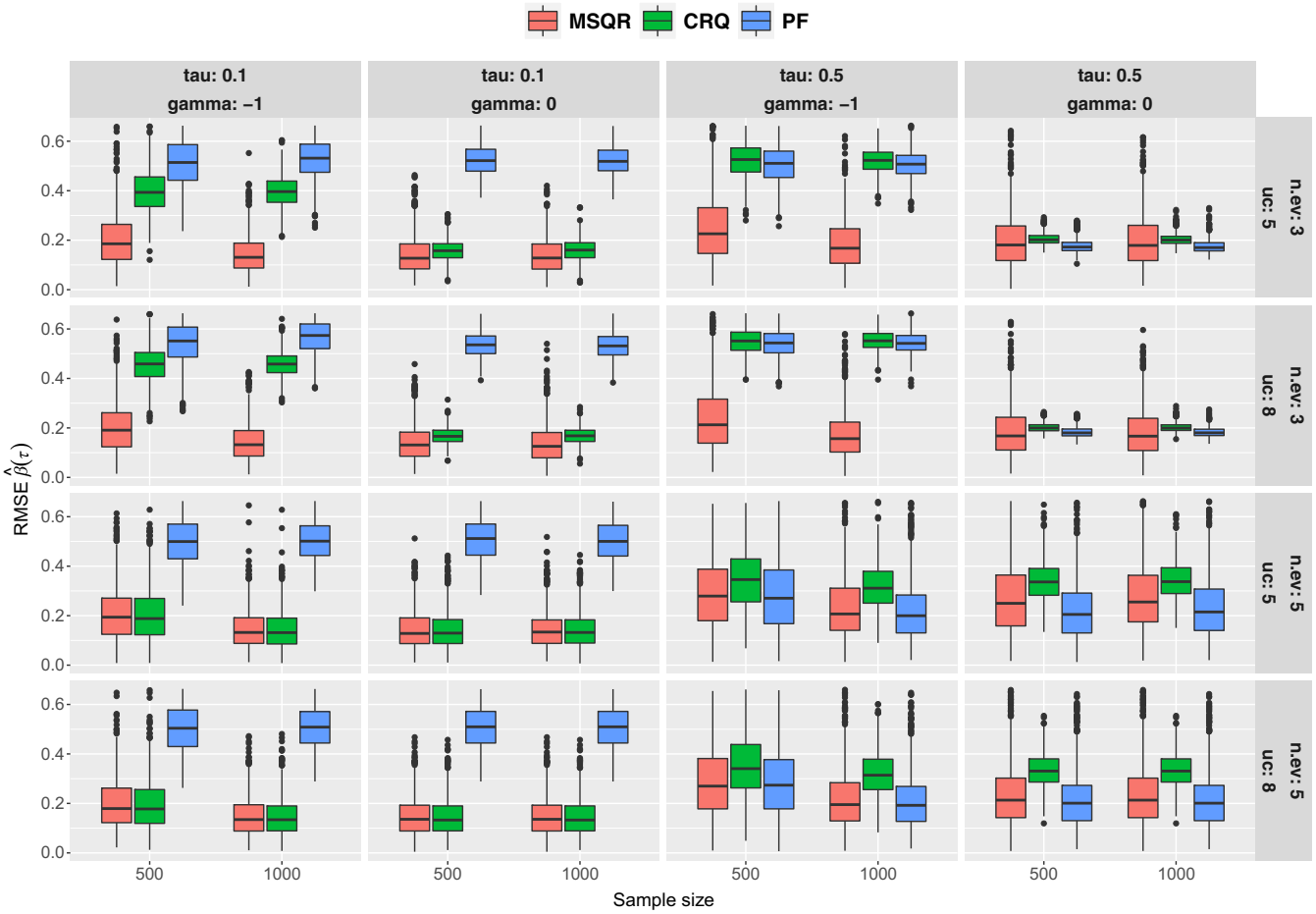

FIGURE 1 Root mean squared error (RMSE) for $\widehat{\beta}(\tau)$ when data are generated according to a competing risk scenario and times are log-normal. Boxplots are based on $B = 1000$ replicates. Multistate quantile regression (MSQR): our proposed approach. Censored quantile regression (CRQ): standard censored quantile regression.[12] PF: competing risk quantile regression cmprskQR[21] [Colour figure can be viewed at wileyonlinelibrary.com]

and Fine even when observations are genuinely competing risk data. We speculate that this is partly consequence of the larger variance associated with inverse probability weighting,[21] which is a crucial ingredient in CRQ. However, we note that, in several cases, MSQR has an advantage also in terms of bias. In a few cases, CRQ has a performance similar to that of MSQR. This can be explained by the setup of our simulation study, in which coefficients and predictions for different transitions coincide, thus increasing the effective sample size available to the CRQ estimator that does not discriminate transitions based on state of origin. Our conclusions also extend to the results for the RMSE and bias of $\widehat{Q}(\tau)$ when times are log-normal, as well as to those obtained when times are generated using Weibull distributions (see supplementary material).

4 | APPLICATION TO CROSS-INFECTIONS IN CIRRHOTIC PATIENTS

In this section, we analyze data from an observational study on $n = 870$ cirrhotic patients who were admitted to the gastroenterology department of a public hospital ("Policlinico Umberto I") in Rome, Italy, between October 2008 and March 2017. The main aim of the study was to assess both risk and protective factors of two outcomes: occurrence of nosocomial infections (primary prevention) and mortality after infection (secondary prevention). Standard precautions for preventing spread of hospital infections include hand hygiene before and after every patient contact, use of gloves, gowns, and eye protection (for situations in which exposure to body fluids is possible). The available information includes baseline characteristics of the patients and the procedures they underwent during the hospital stay. For our analysis, we selected age at first admission, gender, catheterization during hospital stay, history of infections, model for end-stage liver disease (MELD) score, paracentesis during hospital stay, and alcohol abuse. For patients with an infection at

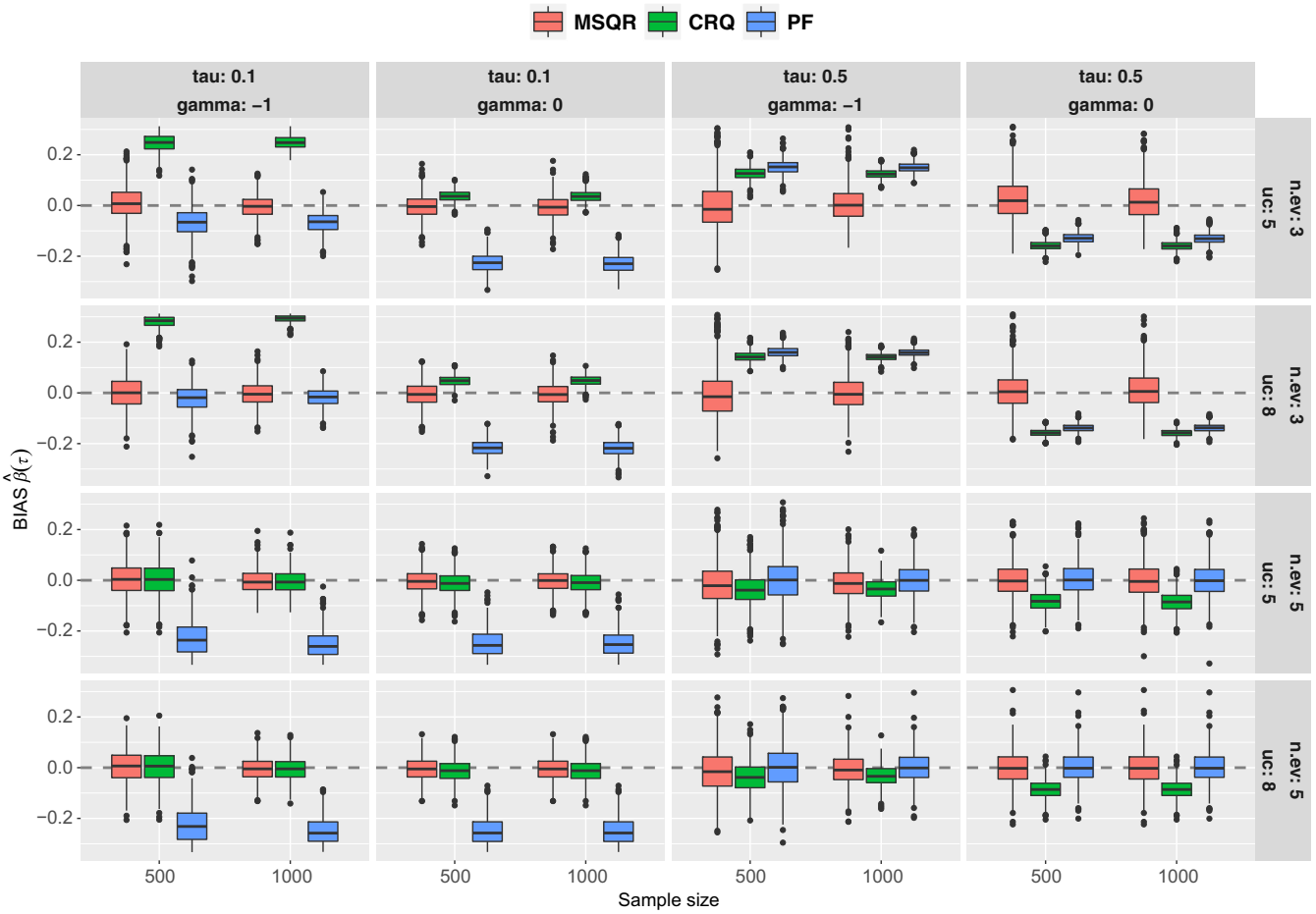

FIGURE 2 Bias for $\widehat{\beta}(\tau)$ when data are generated according to a competing risk scenario and times are log-normal. Boxplots are based on $B = 1000$ replicates. Multistate quantile regression (MSQR): our proposed approach. Censored quantile regression (CRQ): standard censored quantile regression.[12] PF: competing risk quantile regression cmprskQR[21] [Colour figure can be viewed at wileyonlinelibrary.com]

admission or developed during hospital stay, we distinguish between hospital-acquired (HA) infections and community/health care-acquired (CA/HCA) infections.

We cast the data into a multistate framework by defining four possible states indexed, respectively, from 1 to 4: noninfected, HA infected, CA/HCA infected, and dead. The distinction between CA/HCA and HA infections is important for two reasons. First, HA infections are known to be associated with increased risk of morbidity and mortality, and to be related to infectious agents that are often resistant to antibiotics. In this regard, Italy is considered a country at high risk of HA infections and of multiple antibiotic resistance. The other reason for the distinction between types of infections is that HA infections are a proxy for quality of hospital care, being connected to hospital admission.[41]

The transition matrix of the four possible states is reported in Table 3. The absence of censoring in transitions from infected states is due to the fact that patients were not discharged until the infection had resolved.

Mean age and standard deviation were 60.9 and 11.87 years, respectively, with a male-to-female ratio of about 3:1. Approximately 37% of the patients underwent a paracentesis, 17% a catheterization, and 9% both procedures. A history of infection was present in 25% of the patients, while 20% had a history of alcohol abuse. Median MELD was 13, with 35% of the patients having a value above 15 (the cut-off used for admission to the waiting list for liver transplantation).

Since some states are transient, QR methods developed for the analysis of standard survival or competing risk data are not appropriate. Additionally, since several patients have only one transition, classical MSMs have an unbounded likelihood and therefore are not estimable.

Our aim is to model shorter and longer times to transition to infection or death using catheterization, paracentesis, and alcohol abuse as main exposures, after adjusting for age, gender, and history of infections. We are particularly interested in transitions to HA infected, and from HA infected to death. However, in our analysis, the state of origin for transitions to HA infected is irrelevant. Therefore, we merged states 1 and 3 into a single state, assuming homogeneity of coefficients.

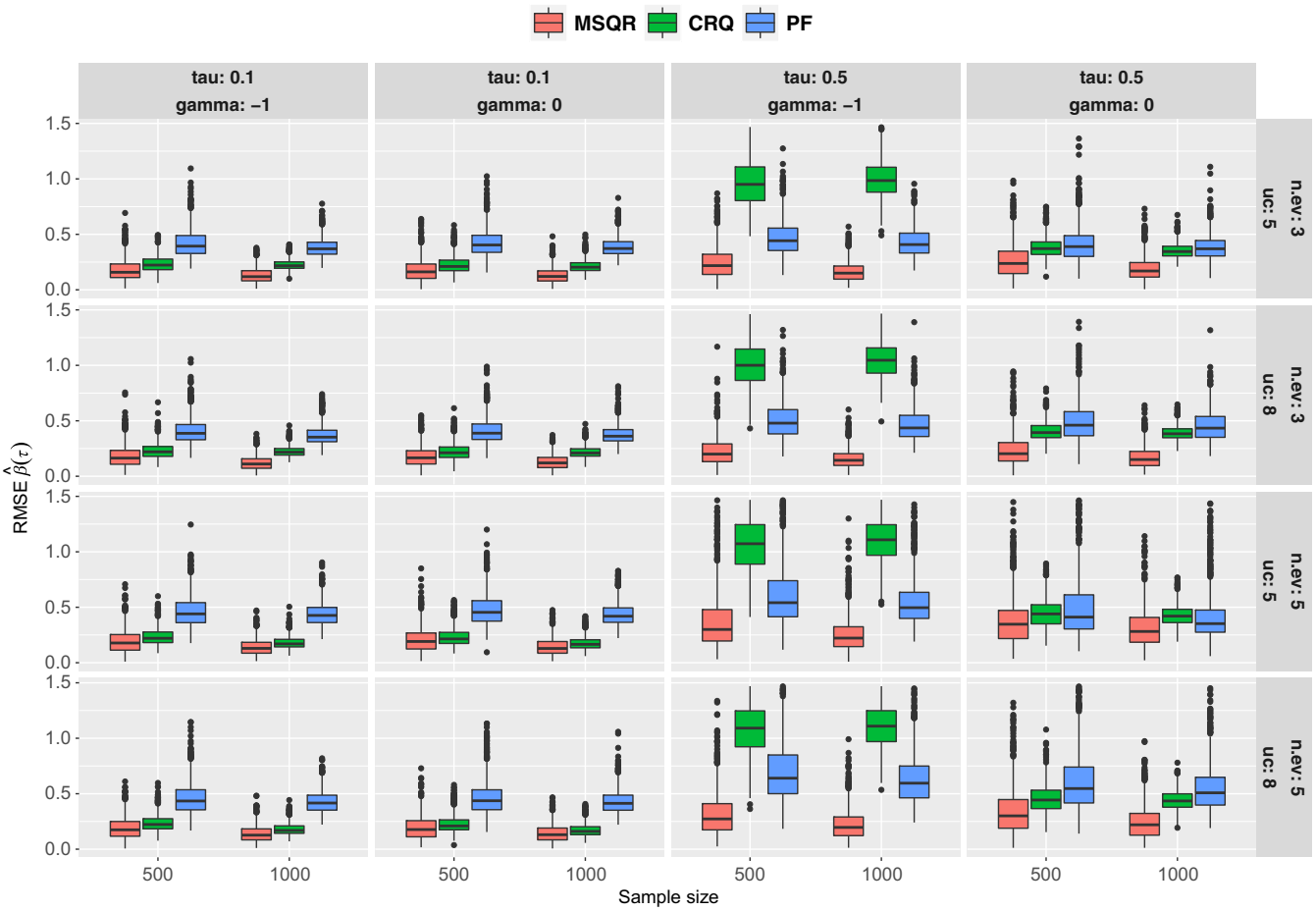

FIGURE 3 Root mean squared error (RMSE) for $\widehat{\beta}(\tau)$ when data are generated according to a multistate scenario and times are log-normal. Boxplots are based on $B = 1000$ replicates. Multistate quantile regression (MSQR): our proposed approach. Censored quantile regression (CRQ): standard censored quantile regression.[12] PF: competing risk quantile regression cmprskQR[21] [Colour figure can be viewed at wileyonlinelibrary.com]

We then estimated MSQR models for 16 quantiles at level $\tau \in \{i/100 : i = 3, \dots, 18\}$ for transition to HA infected, and 8 quantiles at level $\tau \in \{i/100 : i = 3, \dots, 10\}$ for transition to death. These ranges of τ values were defined following an evaluation of estimable quantiles as discussed in Section 2.1. We work with time on a logarithmic scale ($g_{mj}(\cdot) = \log(\cdot)$), which we have found to lead to more stable results with respect to the identity transformation. Time is expressed in days.

Before we can proceed with the discussion of the modeling results, we need to verify whether the homogeneity assumption is supported by the data. In Figure 5, we report the p-values of the test described in Section 2.2. We conclude that the data support our homogeneity assumption at the 5% significance level across all considered quantiles.

The estimated coefficients along with 95% bootstrapped confidence intervals are shown in Figure 6. There are significant effects of catheterization on transitions to HA infected state. Indeed, this procedure is notoriously associated with a substantial risk of HA infections. However, our analysis shows that the magnitude of the effect is larger at higher quantiles. This means that, as compared to patients without catheter with longer permanence in transient states, those with catheter experience a much shorter permanence. The negative effects of paracentesis, too, show a magnitude that increases with quantile level, although they are not significant at the 5% level. Neither catheterization nor paracentesis are significantly associated with time to death.

Finally, alcohol abuse does not seem to have a significant effect, neither practically nor statistically, on any of the quantiles of the time-to-event distribution of progression to either HA infection or death. This is explained by the presence in the model of direct measures of progression to hepatic failure (which is known to be associated with impaired immune response) like MELD and history of infections, and it is confirmed by models estimated without confounders.

We conclude this section by showing the same results in the familiar form of adjusted survival estimates. The broken lines in Figure 7 were obtained by connecting points with coordinates $\widehat{Q}(\tau \mid Z)$ and $(1 - \tau)$. Predictions $\widehat{Q}(\tau \mid Z)$ were

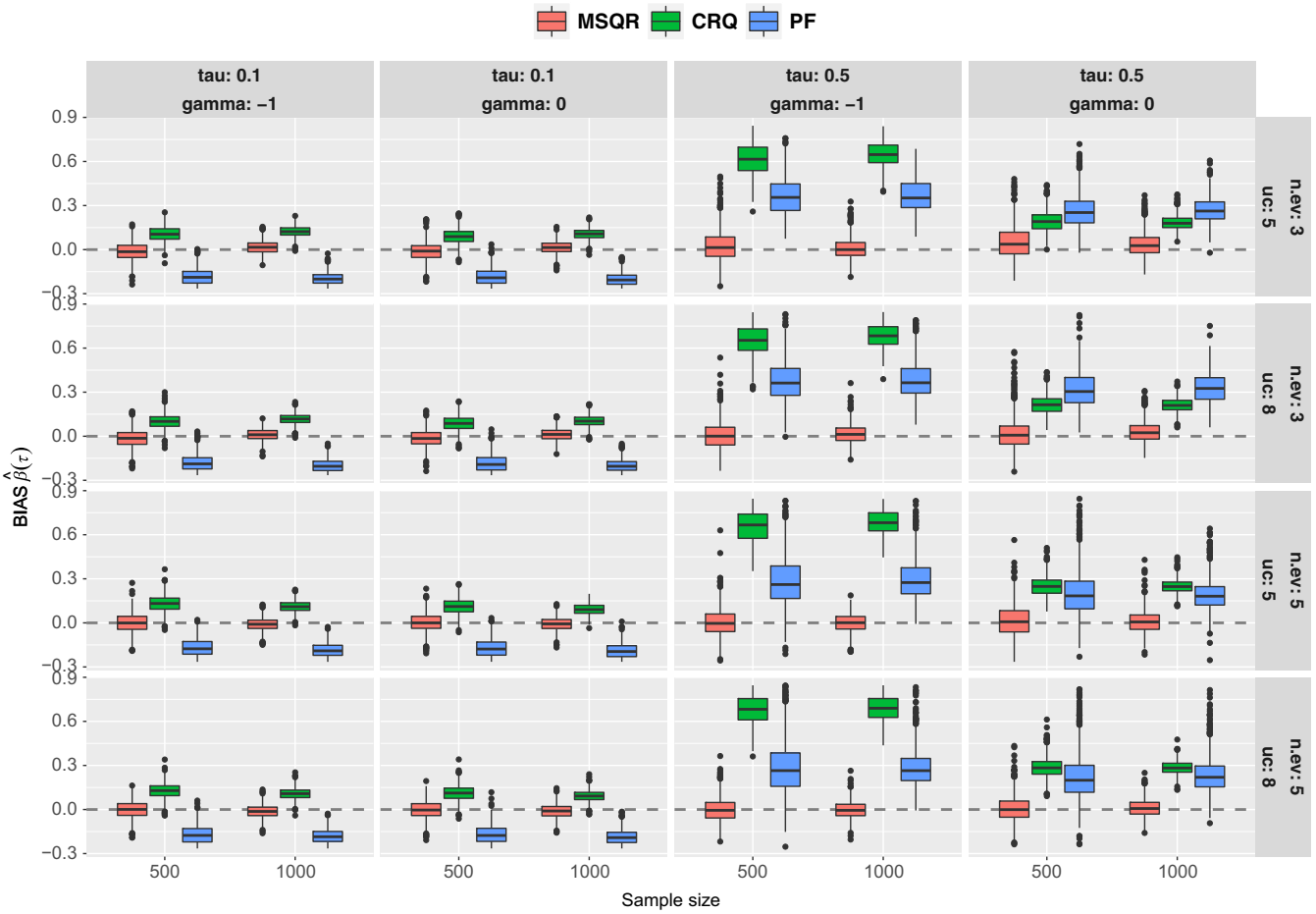

FIGURE 4 Bias for $\widehat{\beta}(\tau)$ when data are generated according to a multistate risk scenario and times are log-normal. Boxplots are based on $B = 1000$ replicates. Multistate quantile regression (MSQR): our proposed approach. Censored quantile regression (CRQ): standard censored quantile regression.[12] PF: competing risk quantile regression cmprskQR[21] [Colour figure can be viewed at wileyonlinelibrary.com]

	Censored	Not infected	HA infected	CA/HCA infected	Dead
Not infected	573	0	124	15	9
HA infected	0	101	0	0	50
CA/HCA infected	0	156	27	0	20

Abbreviations: CA/HCA, community/health care-acquired; HA, hospital-acquired.

TABLE 3 Transition matrix for four possible states in the cirrhotic patients data. Repeated events are counted

obtained for the "most likely" subject in the dataset: a 65-year-old male with no history of infections. A word of caution regarding quantile crossing is in order. Since quantiles are estimated separately, monotonicity of $Q(\tau \mid Z)$ is not guaranteed. This does not occur in our data, but it could certainly happen in general. In that case, a solution is, for instance, to rearrange the fitted values.[42]

5 | FINAL REMARKS

The statistical analysis of disease histories is fundamental for the assessment of etiology and prognosis, efficacy and cost effectiveness of treatments, and, in general, quality of life. In this regard, MSMs have gained an important role in clinical and epidemiological studies.

In this paper, we modeled conditional quantiles of times between the occurrence of events. The advantages of our QR models over classical MSMs are manifold. First of all, regression quantiles are interpreted on the scale of the outcome. For example, the negative effect of a covariate corresponds to a shorter permanence in any given state for an increase in that covariate, all else being equal. This is tantamount to a *positive* log-hazard ratio, which, however, does not by itself

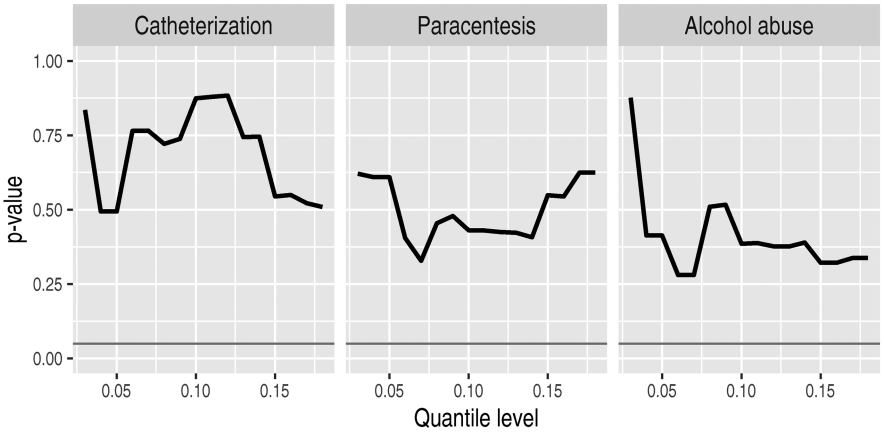

FIGURE 5 P-values of the test on homogeneity of transitions from noninfected or CA/HCA infected to HA infected, shown separately for catheterization, paracentesis, and alcohol abuse. All p-values are adjusted for age, gender, history of infections, and model for end-stage liver disease score. The reference (gray horizontal lines) is drawn at 0.05. CA/HCA, community/health care-acquired; HA, hospital-acquired

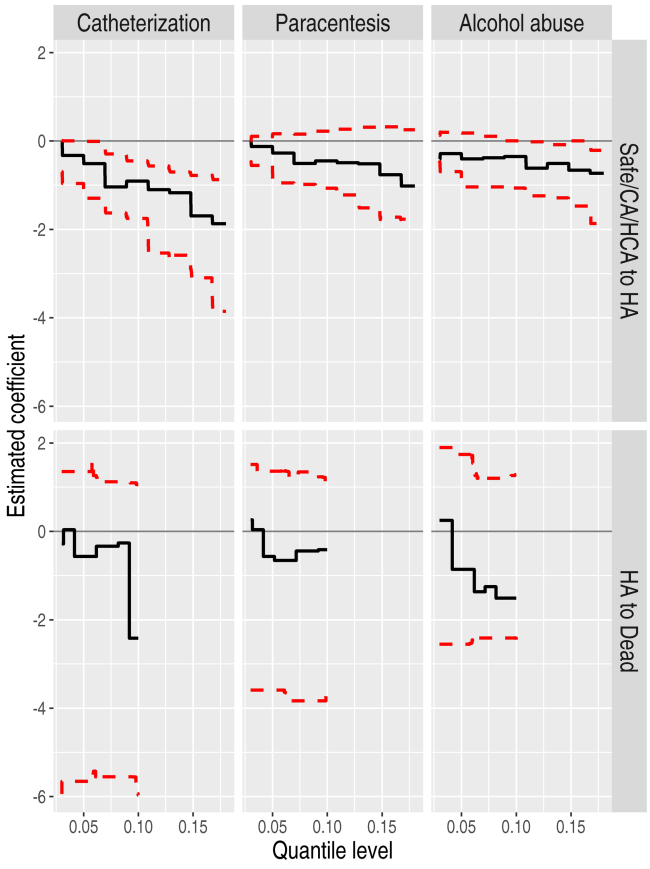

FIGURE 6 Estimated coefficients (black solid lines) and 95% confidence intervals (red dashed lines) at different quantiles, shown separately for catheterization, paracentesis, and alcohol abuse. All estimates are adjusted for age, gender, history of infections, and model for end-stage liver disease score. The reference (gray horizontal lines) is drawn at 0. CA/HCA, community/health care-acquired; HA, hospital-acquired [Colour figure can be viewed at wileyonlinelibrary.com]

provide any information on duration, much less on specific quantiles of the conditional time-to-event distribution. Other advantages of a quantile approach consist in the ability to model survival percentiles without having to make distributional assumptions and robustness of results to the presence of outliers. Moreover, by focusing on selected transitions, which are possible but rare (eg, from not infected to CA/HCA infected), we are still able to harness the available information. In contrast, classical MSMs might be unstable, or even not estimable without regularization, as they aim at estimating the entire transition matrix.

To our knowledge, ours is the first proposal of quantile models with general applicability to multistate processes, encompassing all special cases (eg, competing risks, semicompeting risks, and recurrent events). Our methodology starts from a convenient representation of the data and subsequent application of standard methods for QR with censored data. In a simulation study, the proposed estimator substantially outperformed existing methods, including those devised for competing risk analysis under a competing risk scenario. We also discussed homogeneity constraints, partial homogeneity constraints, and how to test them.

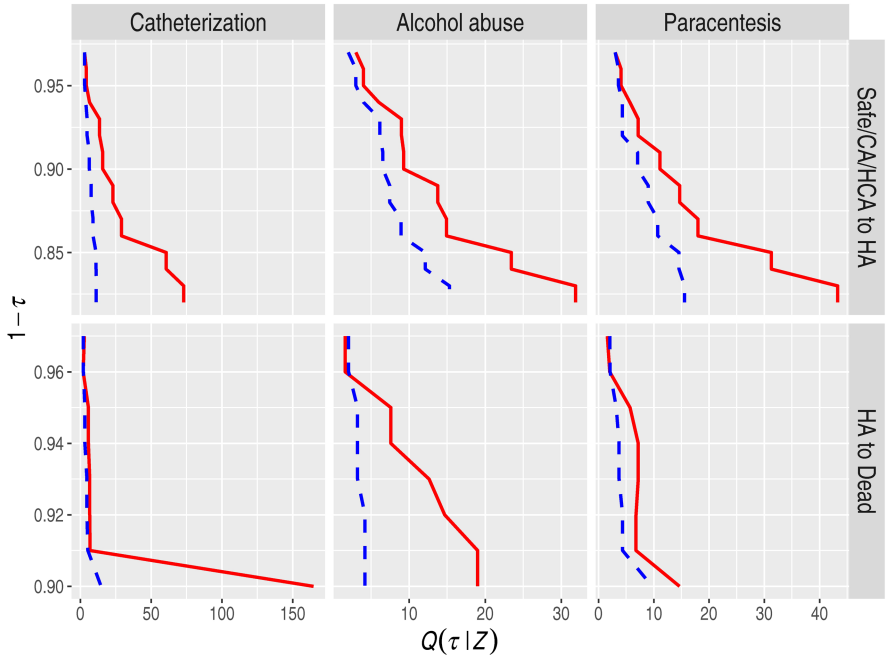

FIGURE 7 Adjusted estimate survival curves contrasting 65-year-old males with no history of infections who either have (dashed blue line) or do not have (continuous red line) one of the risk factors reported by column, for the transitions reported by row. CA/HCA, community/health care-acquired; HA, hospital-acquired [Colour figure can be viewed at wileyonlinelibrary.com]

Finally, we demonstrated the application of the proposed methods to assess time-to-infection and time-to-death in cirrhotic patients. We reported the novel discovery that the association between catheterization and time-to-HA infection is quantile-dependent, with stronger, negative estimates at larger quantiles. That is, shorter times to HA infection are not affected by catheterization as much as longer times. We conclude that catheterization can be a potentially harmful procedure for cirrhotic and/or immunosuppressed patients and might have a causal role in the onset of HA infections.

ACKNOWLEDGEMENTS

The authors are grateful to three anonymous referees for helpful suggestions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

ORCID

Alessio Farcomeni https://orcid.org/0000-0002-7104-5826
Marco Geraci https://orcid.org/0000-0002-6311-8685

REFERENCES

1. Putter H, Fiocco R, Geskus R. Tutorial in biostatistics: competing risks and multi-state models. *Statist Med*. 2007;26:2277-2432.
2. Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJ. Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal*. 2006;48(3):366-80.
3. Meier-Hirmer C, Schumacher M. Multi-state model for studying an intermediate event using time-dependent covariates: application to breast cancer. *BMC Med Res Methodol*. 2013;13:80.
4. Pashayan N, Pharoah P, Neal DE, et al. PSA-detected prostate cancer and the potential for dedifferentiation—estimating the proportion capable of progression. *Int J Cancer*. 2011;128:1462-1470.
5. Álvaro-Meca A, Kneib T, Prieto RG, de Miguel AG. Impact of comorbidities and surgery on health related transitions in pancreatic cancer admissions: a multi state model. *Cancer Epidemiology*. 2012;36:142-146.
6. Álvaro-Meca A, Akerkar R, Alvarez-Bartolome M, Gil-Prieto R, Rue H, de Miguel AG. Factors involved in health-related transitions after curative resection for pancreatic cancer. 10-years experience: a multi state model. *Cancer Epidemiology*. 2013;37:91-96.
7. Iacobelli S, de Wreede LC, Schönland S, et al. Impact of CR before and after allogeneic and autologous transplantation in multiple myeloma: results from the EBMT NMAM2000 prospective trial. *Bone Marrow Transplant*. 2015;50:505-510.

8. Lauseker M, Hasford J, Hoffmann VS, Muller MC, Hehlmann R, Pfirrmann M. A multi-state model approach for prediction in chronic myeloid leukaemia. *Annal Hematology*. 2015;94:919-927.
9. Eulenburg C, Mahner S, Woelber L, Wegscheider K. A systematic model specification procedure for an illness-death model without recovery. *PLOS ONE*. 2015;10(4):e0123489.
10. Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978;46(1):33-50.
11. Koenker R, Geling O. Reappraising medfly longevity: a quantile regression survival analysis. *J Am Stat Assoc*. 2001;96(454):458-468.
12. Portnoy S. Censored regression quantiles. *J Am Stat Assoc*. 2003;98:1001-1012.
13. Peng L, Huang Y. Survival analysis with quantile regression models. *J Am Stat Assoc*. 2008;103:637-649.
14. Yin G, Zeng D, Li H. Power-transformed linear quantile regression with censored data. *J Am Stat Assoc*. 2008;103:1214-1224.
15. Wang HJ, Wang L. Locally weighted censored quantile regression. *J Am Stat Assoc*. 2009;104:1117-1128.
16. Frumento P, Bottai M. An estimating equation for censored and truncated quantile regression. *Comput Stat Data Anal*. 2017;113:53-63.
17. Chrisout E, Akritas MG. Single index quantile regression for censored data. *Stat Method Appl*. 2019;28:655-678.
18. Yang X, Narisetty NN, He X. A new approach to censored quantile regression estimation. *J Comput Graph Stat*. 2018;27:417-425.
19. De Backer M, El Ghouch A, van Keilegom I. An adapted loss function for censored quantile regression. *J Am Stat Assoc*. 2019. https://doi.org/10.1080/01621459.2018.1469996
20. Peng L, Fine JP. Nonparametric quantile inference with competing-risks data. *Biometrika*. 2007;94:735-744.
21. Peng L, Fine JP. Competing risks quantile regression. *J Am Stat Assoc*. 2009;104:1440-1453.
22. Beyersmann J, Schumacher M. A note on nonparametric quantile inference for competing risks and more complex multistate models. *Biometrika*. 2008;95:1006-1008.
23. Li R, Peng L. Quantile regression for left-truncated semicompeting risks data. *Biometrics*. 2011;67:701-710.
24. Sun X, Peng L, Huang Y, Lai HJ. Generalizing quantile regression for counting processes with applications to recurrent events. *J Am Stat Assoc*. 2016;111:145-156.
25. Hsieh J-J, Wang H-R. Quantile regression based on counting process approach under semi-competing risks data. *Ann Inst Stat Math*. 2018;70:395-419.
26. Ma H, Peng L, Huang C-Y, Fu H. Quantile regression modeling of recurrent event risk. ArXiv e-prints. 2018. https://arxiv.org/abs/1811.06211
27. Meira-Machado L, de Uña Alvarez J, Cadarso-Suarez C, Andersen PK. Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res*. 2009;18:195-222.
28. Marino MF, Farcomeni A. Linear quantile regression models for longitudinal experiments: an overview. *METRON*. 2015;73:229-247.
29. Koenker R. Quantile regression for longitudinal data. *J Multivar Anal*. 2004;91:74-89.
30. Kim M-O, Yang Y. Semiparametric approach to a random effects quantile regression model. *J Am Stat Assoc*. 2011;106:1405-1417.
31. Farcomeni A. Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Stat Comput*. 2012;22:141-152.
32. Geraci M, Bottai M. Linear quantile mixed models. *Stat Comput*. 2014;24(3):461-479.
33. Farcomeni A, Viviani S. Longitudinal quantile regression in presence of informative drop-out through longitudinal-survival joint modeling. *Statist Med*. 2015;34:1199-1213.
34. Arellano M, Bonhomme S. Nonlinear panel data estimation via quantile regression. *Econometrics Journal*. 2016;19:61-94.
35. Geraci M. Modelling and estimation of nonlinear quantile regression with clustered data. *Comput Stat Data Anal*. 2019;136:30-46.
36. Karlsson A. Nonlinear quantile regression estimation of longitudinal data. *Commun Stat Simul Comput*. 2008;37:114-131.
37. Leng C, Zhang W. Smoothing combined estimating equations in quantile regression for longitudinal data. *Stat Comput*. 2014;24:123-136.
38. Oberhofer W, Haupt H. Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory*. 2016;32:686-713.
39. Oberhofer W, Haupt H. *Consistency of Nonlinear Regression Quantiles Under Type I Censoring Weak Dependence and General Covariate Design* [dissertation]. Regensburg, Germany: University of Regensburg; 2005.
40. Jackson CH. Multi-state models for panel data: the MSM package for R. *J Stat Softw*. 2011;38.
41. Escolano S, Golmard J-L, Korinek A-M, Mallet A. A multi-state model for evolution of intensive care unit patients: prediction of nosocomial infections and deaths. *Statist Med*. 2000;19:3465-3482.
42. Chernozhukov V, Fernandez-Val I, Galichon A. Quantile and probability curves without crossing. *Econometrica*. 2010;78:1093-1125.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Farcomeni A, Geraci M. Multistate quantile regression models. *Statistics in Medicine*. 2019;1–12. https://doi.org/10.1002/sim.8393