Journal of the Royal Statistical Society Series A: Statistics in Society, 2026, 00, 1–23
https://doi.org/10.1093/jrsssa/qnag070
Published: 2 June 2026
Original article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series A Statistics in Society
OXFORD

Joint modelling of in-hospital mortality and length of stay: a copula additive distributional regression analysis of COVID-19 patient data

Giampiero Marra[1], **Rosalba Radice**[2]

[1]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
[2]Faculty of Actuarial Science and Insurance, Bayes Business School, City St George's, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK
Address for correspondence: Rosalba Radice, Faculty of Actuarial Science and Insurance, Bayes Business School, City St George's, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK. Email: rosalba.radice@citystgeorges.ac.uk

Abstract

In-hospital mortality and length of stay are fundamental metrics for evaluating healthcare quality, patient outcomes, and resource utilization. While length of stay reflects hospital efficiency and capacity management, mortality provides insights into patient safety and the effectiveness of clinical interventions. These outcomes are interdependent, and demographic, clinical and laboratory factors simultaneously influence both hospitalization duration and mortality. To address this, a copula additive distributional regression framework is employed, enabling the joint modelling of these hospital metrics as functions of covariate effects. Application to COVID-19 data demonstrates that key predictors, including age, oxygenation and inflammation markers, modulate the dependence between mortality and hospitalization duration. The joint modelling approach provides a probabilistic, patient-level characterization of the interplay between these indicators, supporting risk stratification, resource planning and actionable clinical decision-making.

Keywords copula regression, COVID-19, discharge, hospital length of stay, mortality, patient-level risk

1 Introduction

In-hospital mortality and length of stay are two critical metrics in healthcare evaluation, widely used to assess care quality, hospital performance and patient outcomes (e.g. Han et al., 2022; Lingsma et al., 2018; Stone et al., 2022; Wilder et al., 2022; Wu et al., 2022). Length of stay serves as a key indicator of resource utilization and hospital efficiency, influencing bed availability, staffing needs and financial costs. The Organisation for Economic Co-operation and Development notes that shorter stays reduce costs and optimize the use of resources (Organisation for Economic Co-operation and Development, 2024). Similarly, the Institute for Healthcare Improvement identifies length of stay as a well-established measure of hospital efficiency, directly impacting capacity and expenditures (Institute for Healthcare Improvement, 2024). Mortality, on the other hand, is a fundamental measure of patient safety and the effectiveness of clinical interventions, reflecting both patient risk factors and

Received: October 16, 2025. **Revised:** May 14, 2026. **Accepted:** May 14, 2026

broader systemic healthcare performance. The World Health Organization reports that approximately one in every ten patients is harmed in healthcare settings, with over three million deaths annually attributed to unsafe care (World Health Organization, 2024). This highlights the need to monitor mortality rates closely to enhance patient safety and improve the quality of interventions.

The relationship between mortality and hospitalization duration is sophisticated, shaped by various aspects. Patients with severe conditions or multiple comorbidities often require extended hospital stays due to the complexity of their medical needs. A study on subjects with chronic disabilities found that higher comorbidity is associated with both longer stays and increased mortality rates (Rochon et al., 1996). Conversely, certain high-risk individuals may deteriorate rapidly, leading to early mortality before an extended length of stay occurs. Research shows that older patients, particularly those with multiple chronic conditions, experience significantly longer durations and higher mortality rates than younger, healthier individuals (Lee & Park, 2025). Hospital practices also play a crucial role in influencing both metrics. For instance, the intensity of care provided during the early stages of hospitalization can impact patient outcomes. A study on staffing levels found that increased nursing hours are associated with a reduced risk of mortality, highlighting the importance of adequate staffing in improving discharge rates (Griffiths et al., 2019). Efficient discharge planning is another key factor in optimizing hospitalization duration without compromising patient outcomes, with a meta-analysis demonstrating a significant reduction in length of stay following such interventions (Hunt-O'Connor et al., 2021). Moreover, individualized plans have been linked to shorter stays and lower readmission risks (Kim & Covey, 2022). Delays in discharge planning can lead to prolonged hospitalizations and an increased risk of adverse events. Yen et al. (2022) found that subjects experiencing unplanned readmissions after discharge have a higher likelihood of mortality, stressing the importance of continuous and effective planning services.

This article focuses on in-hospital mortality and length of stay among COVID-19 patients, emphasizing their joint distribution. These outcomes are interconnected, and demographic, clinical, and laboratory factors simultaneously influence both hospitalization duration and mortality. Flexible predictive models have recently been applied to address this issue. For example, Chen et al. (2023) proposed a nonlinear neural network model to jointly predict length of stay and in-hospital mortality from tabular electronic health records. While their approach achieves solid predictive performance, it ultimately treats the two outcomes as conditionally independent given covariates. As a result, relevant aspects of the joint distribution remain unexplored. To address this, the copula additive distributional regression framework of Marra and Radice (2025) is employed. To our knowledge, this is the first study to simultaneously model these hospital metrics, providing a comprehensive probabilistic characterization of them at the individual level. Model estimation is carried out using a computationally efficient and stable penalized maximum likelihood technique, while inference is based on results for models fitted via penalized log-likelihood methods. The framework is implemented in the `R` package `GJRM` (Marra & Radice, 2026), which provides tools for fitting flexible copula regression models and producing intuitive numerical and visual summaries. This makes the approach accessible to healthcare practitioners, facilitating data-driven decision-making and enhancing clinical insights.

The copula regression framework provides a detailed understanding of the complex, intertwined dynamics of in-hospital mortality and length of stay. The proposed analysis of patients admitted with COVID-19 during the first surge of the pandemic in New York City reveals that the probability of discharge and hospitalization duration were positively associated, with stronger dependence between lower likelihood of discharge and shorter lengths of stay. Patient characteristics, such as age, oxygenation, renal function and inflammation markers, modify this dependence. The methodology enables clinicians and healthcare planners to assess patient-level risks in a probabilistic and interpretable way, identify opportunities for targeted intervention and understand how different patient subgroups are likely to progress.

The article is structured as follows. Section 2 introduces the building blocks of the adopted model, while Sections 3 and 4 discuss parameter estimation, selected model-based statistics and inferential aspects. Section 5 presents a detailed case study of COVID-19 patient data, jointly analysing in-hospital mortality and length of stay, and highlighting the practical implications for hospital

operations. The analysis illustrates the complex dynamics of these metrics and the clinically meaningful insights they provide for patient management, risk stratification and resource planning. Finally, Section 6 summarizes the main contributions and outlines directions for future research.

2 The model

Consider a pair of random variables for representing mortality and length of stay: $Y_1 \in \{0, 1\}$ and $Y_2 \in \mathbb{N}_0$. The mortality indicator follows a Bernoulli distribution, $Y_1 \sim \text{Bernoulli}(\mu_1)$, while the length of stay, $Y_2 \sim D_2(\mu_2, \sigma_2)$, is modelled using any of the distributions reported in Table 1. The distributional parameters are specified as $g_{\mu_1}(\mu_1) = \eta_{\mu_1}(\boldsymbol{x}_{\mu_1}; \boldsymbol{\beta}_{\mu_1})$, $\log(\mu_2) = \eta_{\mu_2}(\boldsymbol{x}_{\mu_2}; \boldsymbol{\beta}_{\mu_2})$ and $\log(\sigma_2) = \eta_{\sigma_2}(\boldsymbol{x}_{\sigma_2}; \boldsymbol{\beta}_{\sigma_2})$. Three options are available for $g_{\mu_1}(\cdot)$, each ensuring that $\mu_1 \in (0, 1)$: $g_{\mu_1}(\mu_1) = \Phi^{-1}(\mu_1)$, the inverse of $\Phi(\cdot)$, the cumulative distribution function (CDF) of the standard Gaussian distribution; $g_{\mu_1}(\mu_1) = \log(\mu_1/(1-\mu_1))$, the inverse of the standard logistic CDF; $g_{\mu_1}(\mu_1) = \log(-\log(1-\mu_1))$, the inverse of the standard Gumbel CDF. These correspond to the `probit`, `logit` and `cloglog` link functions.

The joint distribution of (Y_1, Y_2) can be represented as

$$\mathbb{P}(Y_1 = 0, Y_2 \le y_2) = C(F_1(0; \mu_1), F_2(y_2; \mu_2, \sigma_2); \theta), \tag{1}$$

where $F_1(0; \mu_1) = \mathbb{P}(Y_1 = 0)$, $F_2(y_2; \mu_2, \sigma_2)$ denotes the marginal CDF of Y_2, $C : (0, 1)^2 \to (0, 1)$ is a two-place copula function with dependence parameter θ specified as $g_\theta(\theta) = \eta_\theta(\boldsymbol{x}_\theta; \boldsymbol{\beta}_\theta)$, and $g_\theta(\cdot)$ is a known monotonic one-to-one transformation ensuring that θ remains within its valid range. Table 2 presents the available choices for specifying the copula. These families capture a wide spectrum of dependence patterns, including diverse tail behaviours and strengths of association, offering the flexibility needed for a broad range of applications; see Joe (2014) and Nelsen (2006) for a comprehensive introduction to copula theory. For copulae that only support positive dependence (e.g. Clayton and Joe), counter-clockwise rotated versions are obtained as follows: $C_{90}(u_1, u_2; \theta) = u_2 - C(1-u_1, u_2; \theta)$, $C_{180}(u_1, u_2; \theta) = u_1 + u_2 - 1 + C(1-u_1, 1-u_2; \theta)$ and $C_{270}(u_1, u_2; \theta) = u_1 - C(u_1, 1-u_2; \theta)$, where the subscript of C indicates the degree of rotation, and u_1 and u_2 are the shorthand notations for the marginal CDFs used in equation (1). The additive predictor $\eta_\cdot(\boldsymbol{x}_\cdot; \boldsymbol{\beta}_\cdot) \in \mathbb{R}$ depends on a set of regressors $\boldsymbol{x}_\cdot$ and parameter vector $\boldsymbol{\beta}_\cdot$, allowing for various types of covariate effects as detailed in Section 2.1.

Recalling that the marginal probability mass function (PMF) of Y_2 can be expressed as $f_2(y_2; \mu_2, \sigma_2) = F_2(y_2; \mu_2, \sigma_2) - F_2(y_2 - 1; \mu_2, \sigma_2)$, the joint PMF of (Y_1, Y_2) is given by

$$f_{12}(y_1, y_2; \mu_1, \mu_2, \sigma_2, \theta) = Q_1^{1-y_1} Q_2^{y_1}, \tag{2}$$

Table 1. Definition and key properties of selected count distributions

Distribution	$f(y; \mu, \sigma)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$
Poisson (`P`)	$\frac{\exp(-\mu)\mu^y}{y!}$	μ	μ
Negative binomial type I (`NBI`)	$\frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$	μ	$\mu + \sigma\mu^2$
Negative binomial type II (`NBII`)	$\frac{\Gamma(y+\mu/\sigma)\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1+\sigma)^{y+\mu/\sigma}}$	μ	$(1+\sigma)\mu$
Poisson Inverse Gaussian (`PIG`)	$\left(\frac{2\varpi}{\pi}\right)^{0.5}\frac{\mu^y \exp(1/\sigma)\Upsilon_{y-0.5}(\varpi)}{(\varpi\sigma)^y y!}$	μ	$\mu + \sigma\mu^2$

Note. The distributional parameters μ and σ take values in $(0, \infty)$, while $y \in \mathbb{N}_0$. Since the parameters must be positive, the transformation function $g_\cdot(\cdot) = \log(\cdot)$ is applied in all cases. $\Gamma(\cdot)$ is the gamma function, $\varpi = \sqrt{\frac{1}{\sigma^2} + \frac{2\mu}{\sigma}}$ and $\Upsilon_\hbar(\varpi) = \frac{1}{2}\int_0^\infty x^{\hbar-1}\exp\{-0.5\varpi(x + x^{-1})\}\,dx$ is the modified Bessel function of the third kind. These distributions are parametrized as in Stasinopoulos et al. (2017), to which the reader is referred for details.

Table 2. Definition of various copulae, along with the corresponding parameter range for θ and one-to-one transformation function of θ

Copula	$C(u_1, u_2; \theta)$	Range of θ	$g_\theta(\theta)$
Ali-Mikhail-Haq (AMH)	$\frac{u_1 u_2}{1-\theta(1-u_1)(1-u_2)}$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Clayton (C0)	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$(0, \infty)$	$\log(\theta)$
Farlie-Gumbel-Morgenstern (FGM)	$u_1 u_2\{1 + \theta(1-u_1)(1-u_2)\}$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Frank (F)	$-\theta^{-1}\log\{1 + (\exp\{-\theta u_1\} - 1)$ $(\exp\{-\theta u_2\} - 1)/(\exp\{-\theta\} - 1)\}$	$\mathbb{R}\backslash\{0\}$	–
Galambos (GAL0)	$u_1 u_2 \exp[\{(-\log u_1)^{-\theta}$ $+(-\log u_2)^{-\theta}\}^{-1/\theta}]$	$(0, \infty)$	$\log(\theta)$
Gaussian (N)	$\Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta)$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Gumbel (G0)	$\exp[-\{(-\log u_1)^{\theta}$ $+(-\log u_2)^{\theta}\}^{1/\theta}]$	$[1, \infty)$	$\log(\theta - 1)$
Joe (J0)	$1 - \{(1-u_1)^{\theta} + (1-u_2)^{\theta}$ $-(1-u_1)^{\theta}(1-u_2)^{\theta}\}^{1/\theta}$	$(1, \infty)$	$\log(\theta - 1)$
Plackett (PL)	$(O_1 - \sqrt{O_2})/\{2(\theta - 1)\}$	$(0, \infty)$	$\log(\theta)$
Student's *t* (T)	$t_{2,\varphi}(t_\varphi^{-1}(u_1), t_\varphi^{-1}(u_2); \varphi, \theta)$	$[-1, 1]$	$\tanh^{-1}(\theta)$

Note. Here, u_1 and u_2 are the shorthand notations for the marginal CDFs in equation (1), $\Phi_2(\cdot, \cdot; \theta)$ denotes the CDF of the standard bivariate Gaussian distribution with correlation coefficient θ, $\Phi(\cdot)$ is the CDF of the standard univariate Gaussian distribution, $t_{2,\varphi}(\cdot, \cdot; \varphi, \theta)$ represents the CDF of the standard bivariate Student-t distribution with correlation θ and $\varphi \in (2, \infty)$ degrees of freedom, $t_\varphi(\cdot)$ is the CDF of the standard univariate Student-t distribution with φ degrees of freedom, $O_1 = 1 + (\theta - 1)(u_1 + u_2)$ and $O_2 = O_1^2 - 4\theta(\theta - 1)u_1 u_2$.

where Q_1 and Q_2 denote the probability contributions for Y_2 associated with $Y_1 = 0$ and $Y_1 = 1$, respectively. Specifically, $Q_1 = C(F_1(0; \mu_1), F_2(y_2; \mu_2, \sigma_2); \theta) - C(F_1(0; \mu_1), F_2(y_2 - 1; \mu_2, \sigma_2); \theta)$, and $Q_2 = f_2(y_2; \mu_2, \sigma_2) - Q_1$.

The main practical advantage of copulae is that they allow the construction of a multivariate distribution from arbitrary marginal CDFs and a specified dependence function. Another key benefit is that the selection of the marginal distributions and the dependence structure can be treated as separate but related tasks, which aids in model building. A potential challenge arises when the margins are not continuous, as this can affect the identifiability of the copula function. However, as noted by several authors (e.g. Trivedi & Zimmer, 2007; Yang et al., 2020), this issue is generally not a concern in a regression context with one or more continuous covariates: such regressors expand the ranges of $F_1(0; \mu_1)$ and $F_1(y_2; \mu_2, \sigma_2)$ from discrete points to continuous intervals, ensuring the copula is uniquely determined within the region defined by their possible values.

2.1 Additive predictor

For ease of notation, consider an arbitrary η_i. The use of additive predictors provides a flexible modelling framework in which different types of covariate effects can be captured and estimated from the data, without relying on strong a priori assumptions about their functional relationships (Wood, 2017).

An additive predictor can be generically defined as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\boldsymbol{r}_{ki}),$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept and $\boldsymbol{r}_{ki}$ denotes the k^{th} sub-vector of the complete covariate vector $\boldsymbol{r}_i$, obtained as the union of $\boldsymbol{x}_{i\mu_1}$, $\boldsymbol{x}_{i\mu_2}$, $\boldsymbol{x}_{i\sigma_2}$ and $\boldsymbol{x}_{i\theta}$. Each of the K terms is represented as a linear combination of J_k basis functions $b_{kj_k}(\boldsymbol{r}_{ki})$ and regression coefficients $\beta_{kj_k} \in \mathbb{R}$, that is,

$$s_k(\boldsymbol{r}_{ki}) = \sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(\boldsymbol{r}_{ki}).$$

The vector of evaluations $\{s_k(\boldsymbol{r}_{k1}), \ldots, s_k(\boldsymbol{r}_{kn})\}^T$ can be written compactly as $\boldsymbol{R}_k\boldsymbol{\beta}_k$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ_k})^T$ and the design matrix $\boldsymbol{R}_k$ has entries $\boldsymbol{R}_k[i, j_k] = b_{kj_k}(\boldsymbol{r}_{ki})$.

Each $s_k(\cdot)$ term is subject to centering constraints, imposed following the approach of Wood (2017). To control the structural properties of the k^{th} function, such as smoothness, a quadratic penalty of the form $\lambda_k\boldsymbol{\beta}_k^T\mathbf{S}_k\boldsymbol{\beta}_k$ is employed in model fitting, where $\lambda_k \in (0, \infty)$ is the smoothing parameter regulating the trade-off between model fit and parsimony and $\mathbf{S}_k$ depends solely on the chosen spline basis. The overall penalty can be expressed as $\boldsymbol{\beta}^T\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, $\mathbf{S}_{\boldsymbol{\lambda}} = 0 \oplus \lambda_1\mathbf{S}_1 \oplus \cdots \oplus \lambda_K\mathbf{S}_K$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$, with $\oplus$ denoting the direct sum operator.

This formulation accommodates a wide variety of covariate effects, including nonlinear, spatial (e.g. Markov random field) and smooth interaction terms. Multiple types of basis functions and penalty structures are supported by the GJRM package (Marra & Radice, 2026), which builds upon the framework of Wood (2017), to whom the reader is referred for further methodological details. The following sections outline the model terms used in the case study.

2.1.1 Effects of binary and factor variables

In such cases, $s_k(\boldsymbol{r}_{ki}) = \boldsymbol{r}_{ki}^T\boldsymbol{\beta}_k$, where the design matrix $\boldsymbol{R}_k$ is obtained by stacking all covariate vectors $\boldsymbol{r}_{ki}$. These effects are typically unpenalized, so that $\mathbf{S}_k = \mathbf{0}$.

2.1.2 Nonlinear effects

These effects involve continuous covariates, such as age, and can be flexibly estimated from the data using the widely adopted penalized regression spline approach. The primary assumption is global smoothness, which requires the functions to be sufficiently differentiable. For a continuous variable r_{ki}, the design matrix $\boldsymbol{R}_k$ contains the evaluations of the J_k known spline bases $b_{kj_k}(r_{ki})$ for each observation i. To enforce smoothness, a conventional and theoretically sound choice is $\mathbf{S}_k = \int \mathbf{m}_k(r_k)\mathbf{m}_k(r_k)^T dr_k$, where the j_k^{th} element of $\mathbf{m}_k(r_k)$ is given by $\partial^2 b_{kj_k}(r_k)/\partial r_k^2$, and the integration is performed over the range of r_k. This framework accommodates various definitions of basis functions and penalty structures, including penalized cubic regression and B-splines.

When setting up a smooth term, one must specify the type of spline, the number of basis functions J_k and, in most cases, the knot locations. For a one-dimensional smooth, the specific choice of spline basis generally has little impact on the results. Moreover, J_k is usually set to 10, which provides sufficient flexibility in most applications; however, analyses with larger values can be conducted to assess the sensitivity of the estimates to J_k. Knots can be placed evenly across the range of the covariate or according to its percentiles. For thin-plate regression splines, as adopted in the case study, only the choice of J_k is required (Wood, 2017).

3 Estimation

Given a random sample $(y_{i1}, y_{i2}, \boldsymbol{r}_i)_{i=1}^n$, the parameter estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\mu_1}^T, \hat{\boldsymbol{\beta}}_{\mu_2}^T, \hat{\boldsymbol{\beta}}_{\sigma_2}^T, \hat{\boldsymbol{\beta}}_{\theta}^T)^T$ is obtained via penalized maximum likelihood, as described below.

Based on equation (2), the log-likelihood of the mixed binary and count outcomes copula regression model is defined as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (1 - y_{i1}) \log(Q_{i1}) + y_{i1} \log(Q_{i2}),$$

where Q_{i1} and Q_{i2} are the observation-specific versions of Q_1 and Q_2, with all the parameters indexed by i to account for individual outcomes and covariate effects. Due to the flexibility afforded by the

modelling framework in specifying the model equations, the log-likelihood is augmented by a quadratic penalty. That is,

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\beta}, \tag{3}$$

where $\mathbf{S}_{\boldsymbol{\lambda}}$, now defined for all the additive predictors in the model, is given by $\mathbf{S}_{\boldsymbol{\lambda}\boldsymbol{\beta}_{\mu_1}} \oplus \mathbf{S}_{\boldsymbol{\lambda}\boldsymbol{\beta}_{\mu_2}} \oplus \mathbf{S}_{\boldsymbol{\lambda}\boldsymbol{\beta}_{\sigma_2}} \oplus \mathbf{S}_{\boldsymbol{\lambda}\boldsymbol{\beta}_{\theta}}$, and $\boldsymbol{\lambda}$ collects all the associated smoothing parameter vectors.

Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is performed using the efficient and stable penalized likelihood approach described in Marra et al. (2020), which employs a trust-region algorithm with integrated automatic multiple smoothing parameter selection. When provided with the analytical score and Hessian, the trust-region method converges super-linearly to a point satisfying the second-order sufficient conditions. This approach performs well even for nonconcave problems or functions with nearly flat regions, and is generally more stable and faster than in-line search methods (Nocedal & Wright, 2006, Chapter 4). The computational framework used for the estimation of smoothing parameters also requires the availability of analytical first- and second-order derivatives.

The effective degrees of freedom (*edf*) of a model whose parameters are subject to penalization is given by $edf = \text{tr}[-\mathbf{H}(\hat{\boldsymbol{\beta}})\{-\mathbf{H}_p(\hat{\boldsymbol{\beta}})\}^{-1}]$, where $\text{tr}(\cdot)$ denotes the trace operator, $\hat{\boldsymbol{\beta}}$ is the estimated parameter vector, $\mathbf{H}(\hat{\boldsymbol{\beta}})$ is the Hessian of the negative log-likelihood at $\hat{\boldsymbol{\beta}}$, and $\mathbf{H}_p(\hat{\boldsymbol{\beta}}) = \mathbf{H}(\hat{\boldsymbol{\beta}}) - \mathbf{S}_{\boldsymbol{\lambda}}$ is the penalized Hessian (e.g. Marra & Radice, 2020). Equivalently, $edf = \psi - \text{tr}[\{-\mathbf{H}_p(\hat{\boldsymbol{\beta}})\}^{-1}\mathbf{S}_{\boldsymbol{\lambda}}]$, where $\psi = \dim(\boldsymbol{\beta})$. From this expression, it is clear that as $\boldsymbol{\lambda} \to \mathbf{0}$, $edf \to \psi$, whereas as $\boldsymbol{\lambda} \to \infty$, $edf \to \psi - \zeta$, where ζ is the total number of model parameters subject to penalization. For intermediate values $\mathbf{0} < \boldsymbol{\lambda} < \infty$, the model *edf* lies in the range $[\psi - \zeta, \psi]$. The *edf* of a single smooth or penalized component is obtained by summing the corresponding trace elements.

3.1 Model-based statistics

The copula-based regression framework allows for the direct estimation of the joint PMF of the outcomes for a given patient profile $\tilde{\boldsymbol{r}}$,

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2; \tilde{\boldsymbol{r}}, \boldsymbol{\beta}), \quad y_1 \in \{0, 1\}, \; y_2 \in \mathbb{N}_0.$$

This is crucial for understanding the co-occurrence and dependencies between mortality (or discharge) and length of stay, providing a comprehensive probabilistic characterization of these hospital metrics.

Although not central to the present analysis, natural byproducts of the copula model include expectations, such as $\mathbb{E}(Y_2 \mid Y_1 = 0; \tilde{\boldsymbol{r}}, \boldsymbol{\beta}) = \frac{1}{F_1(0;\, \mu_1)} \sum_{y_2=1}^{\infty} y_2 \, f_{12}(0, y_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \theta)$, as well as conditional probabilities. The infinite sum in the conditional expectations is evaluated numerically by sequentially summing over increasing values of y_2 until the change between successive partial sums becomes negligibly small, for example less than 10^{-5} times the previous sum. The estimators of all these quantities are derived by substituting $\boldsymbol{\beta}$ for $\hat{\boldsymbol{\beta}}$.

4 Inferential aspects

Interval estimation is derived from the framework presented in Wood et al. (2016) for models fitted via penalized log-likelihoods of the general form (3). In particular, inference proceeds under the approximation $\boldsymbol{\beta} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$, with $\mathbf{V}_{\boldsymbol{\beta}} = \{-\mathbf{H}_p(\hat{\boldsymbol{\beta}})\}^{-1}$. The penalty term can be interpreted as favouring smoother model estimates, which is equivalent to imposing a Gaussian prior on $\boldsymbol{\beta}$ of the form $f_{\boldsymbol{\beta}} \propto \exp\{-\boldsymbol{\beta}^T \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\beta}/2\}$. While a Bayesian parameter estimation approach is not adopted here, this connection provides a clear justification for employing $\mathbf{V}_{\boldsymbol{\beta}}$ as the covariance matrix. Notably, this formulation simultaneously accounts for sampling variability and smoothing bias, producing intervals with coverage close to the nominal level (e.g. Marra & Wood, 2012).

For nonlinear functions of the model coefficients, intervals can be conveniently obtained via posterior simulation. For example, a $(1-\vartheta)100\%$ interval for $\mathbb{P}(Y_1 = y_1, Y_2 = y_2; \tilde{\boldsymbol{r}}, \boldsymbol{\beta})$ can be constructed as

follows: draw V random vectors $\boldsymbol{\beta}_v$, $v = 1 \ldots, V$, from the approximate distribution of $\boldsymbol{\beta}$; compute the V corresponding realizations of the function of interest, $\mathbb{P}(Y_1 = y_1, Y_2 = y_2; \tilde{\boldsymbol{r}}, \boldsymbol{\beta}_v)$; take the $(\vartheta/2)$th and $(1 - \vartheta/2)$th empirical quantiles of these values. In practice, ϑ is often set to 0.05, and $V = 100$ generally yields reliable results, although it can be increased if more precision is desired. Note that the resulting distribution of a nonlinear function of the model parameters need not be symmetric.

Well-calibrated p-values for the model terms are obtained following the approach described in Wood (2017, Section 6.12), which employs $\mathbf{V}_{\boldsymbol{\beta}}$ as the covariance matrix.

5 Hospital length of stay and mortality

This study investigates in-hospital length of stay (`los`) and mortality (`death`) among patients admitted with COVID-19 between March 1 and April 16, 2020, modelled jointly through the copula-based regression framework. This period corresponds to the first surge of the COVID-19 pandemic in New York City, when hospitals were experiencing unprecedented case volumes, and clinical management strategies were still rapidly evolving. The dataset was derived from a large urban healthcare system in the Bronx, New York (Montefiore Medical Center and affiliated hospitals), using the Clinical Looking Glass platform (Streamline Health, Atlanta, Georgia) and supplemented with primary medical record review, as originally described in Altschul et al. (2020). The original dataset is publicly available at https://figshare.com/s/79827c396af7df42b3d7?file=24020852.

The study population consists of hospitalized patients with laboratory-confirmed SARS-CoV-2 infection documented at or prior to admission, reflecting predominantly community-acquired infections. Almost all patients tested positive at or before admission, making it unlikely that many acquired COVID-19 during their hospital stay. However, systematic screening for hospital-acquired infections was not performed, so a small number of nosocomial cases cannot be ruled out. Patient characteristics included demographics, comorbidities, vital signs and laboratory values recorded at admission. The variable names and their definitions are summarized in Table 3, while the corresponding descriptive statistics for the study population ($n = 4711$) are presented in Table 4. Several laboratory biomarkers, including D-dimer, glucose, interleukin-6 (IL-6), ferritin and procalcitonin, were excluded from the statistical analysis because a substantial proportion of their values were missing (ranging from approximately 40% to over 70% of observations).

In this cohort, the observed in-hospital mortality rate was approximately 24%, reflecting the high severity of illness and limited treatment options during the first wave of the pandemic (notably higher than mortality rates observed in later waves, when treatment protocols, vaccination and viral variants reduced overall mortality). As the data were collected during the early phase and within a single healthcare network serving a predominantly urban population with a high burden of comorbidities, the models estimated on this dataset may reflect disease severity, clinical practices and mortality risks specific to that period. Therefore, caution is warranted when generalizing findings to later stages of the pandemic or to different healthcare settings.

The analysis was conducted on 3336 complete cases, obtained after excluding observations with missing values. To assess the impact of this exclusion, the outcomes and selected baseline characteristics were compared between the full cohort and the complete-case subset, as reported in Table 5. Length of stay and in-hospital mortality are similar across the two cohorts. The distributions of age and major clinical severity indicators are broadly comparable, suggesting that restricting the analysis to complete cases may not have meaningfully changed the overall composition of the study population. Nevertheless, because missingness may reflect underlying differences in disease severity or clinical care, some degree of selection bias within the complete-case cohort cannot be excluded.

5.1 Model building

As discussed in Section 2, the specification of the marginal models and the dependence structure can be treated as separate but related tasks, which provides a coherent framework for the overall modelling strategy.

Table 3. Definition and coding of the variables used in the case study

R variable name	Description	Coding	Units
los	Length of hospital stay	Discrete variable	days
death	In-hospital mortality	0 = no, 1 = yes	–
mi	History of myocardial infarction	0 = no, 1 = yes	–
pvd	Peripheral vascular disease	0 = no, 1 = yes	–
chf	Congestive heart failure	0 = no, 1 = yes	–
cvd	Cardiovascular disease (other than `mi` or `chf`)	0 = no, 1 = yes	–
dementia	Dementia (any cause)	0 = no, 1 = yes	–
copd	Chronic obstructive pulmonary disease	0 = no, 1 = yes	–
dm.complicated	Diabetes mellitus with complications	0 = no, 1 = yes	–
dm.simple	Diabetes mellitus without complications	0 = no, 1 = yes	–
renal.disease	Chronic kidney disease	0 = no, 1 = yes	–
all.cns	Any central nervous system condition	0 = no, 1 = yes	–
pure.cns	Isolated central nervous system disease	0 = no, 1 = yes	–
stroke	History of stroke	0 = no, 1 = yes	–
seizure	History of seizures	0 = no, 1 = yes	–
old.syncope	History of syncope	0 = no, 1 = yes	–
old.other.neuro	Other neurological history	0 = no, 1 = yes	–
other.brain.lesion	Other brain lesion	0 = no, 1 = yes	–
age	Patient age	Continuous variable	years
map	Mean arterial pressure	Low (<65), Normal (65–100), High (>100)	mmHg
inr	International normalized ratio	Normal (0.7–1.2), Elevated (>1.2)	–
bun	Blood urea nitrogen	Normal ($\leq$ 20), Mildly elevated (21–40), Severely elevated (>40)	mg/dL
creatinine	Serum creatinine	Normal ($\leq$1.2), Mildly elevated (1.3–2.0), High (>2.0)	mg/dL
sodium	Serum sodium	Hyponatremia (<135), Normal (135–145), Hypernatremia (>145)	mmol/L
ast	Aspartate aminotransferase	Normal ($\leq$40), High (>40)	U/L
alt	Alanine aminotransferase	Normal ($\leq$55), High (>55)	U/L
wbc	White blood cell count	Low (<4), Normal (4–11), High (>11)	$\times 10^3/\mu$L

(*continued*)

Table 3. Continued

R variable name	Description	Coding	Units
crp	C-reactive protein	Normal (<5), Mildly elevated (5–20), High (>20)	mg/L
troponin	Serum troponin	Normal (≤0.04), Elevated (>0.04)	ng/mL
race	Race or ethnicity	Asian, Black, Latino, White	–
oxygenation	Oxygen saturation category	Normal (≥95%), Mild hypoxemia (90–94%), Moderate/Severe hypoxemia (≤89%)	%
temperature	Body temperature	Hypothermia (<36°C), Normal (36–37.9°C), Fever (≥38°C)	°C
platelets	Platelet count	Low (<150), Normal (150–450), High (>450)	$\times 10^3/\mu$L
lymphocytes	Lymphocyte count	Low (<1.0), Normal (1.0–3.5), high (>3.5)	$\times 10^3/\mu$L

Table 4. Descriptive statistics of the outcomes, demographics, vital signs, comorbidities and laboratory values in the study population ($n = 4711$)

Outcomes					
los		Median [IQR]			5 [3−9]
		Mean (range)			7.16 (0−56)
death		Proportion			24.4%
Demographics and Vitals					
age	Median [IQR]	65 [54–76]	oxygenation	Normal	55.5%
	Mean (range)	63.4 (18–103)		Mild	24.8%
race	Asian	3.1%		Moderate/Severe	19.7%
	Black	44.4%	temperature	Normal	77.6%
	Latino	42.4%		Fever	20.0%
	White	10.1%		Hypothermia	2.4%
map	Normal	72.0%			
	Low	9.8%			
	High	18.2%			
Comorbidities					
mi	Yes	4.3%	dm.complicated	Yes	10.5%
pvd	Yes	18.0%	dm.simple	Yes	14.6%
chf	Yes	11.5%	renal.disease	Yes	17.7%
cvd	Yes	10.7%	all.cns	Yes	12.9%
dementia	Yes	7.9%	pure.cns	Yes	10.4%
copd	Yes	5.6%	stroke	Yes	1.2%
old.other.neuro	Yes	3.1%	seizure	Yes	0.8%
other.brain.lesion	Yes	0.6%	old.syncope	Yes	1.9%
Laboratory values					
platelets	Normal	78.0%	creatinine	Normal	56.3%
	Low	17.8%		Mild	20.7%
	High	4.2%		High	23.0%
inr	Normal	80.2%	sodium	Normal	62.1%
	Elevated	19.8%		Hyponatremia	27.3%
				Hypernatremia	10.6%
bun	Normal	53.2%	ast	Normal	51.3%
	Mildly elevated	23.9%		High	48.7%
	Severely elevated	22.9%			
alt	Normal	82.0%	wbc	Normal	70.3%
	High	18.0%		Low	7.7%
				High	22.0%
lymphocytes	Normal	52.8%	crp	Normal	30.2%
	Low	45.3%		Mild	48.3%
	High	1.9%		High	21.5%
troponin	Normal	79.4%		Elevated	20.6%

Note. The variables los and age are summarized as median [IQR] and mean (range). For the remaining variables, percentages are shown (absolute counts omitted for compactness). The proportion of missing values in the variables shown in this table ranged from 0% to 19.0%.

Table 5. Comparison of the outcomes and selected predictors between the full cohort and the complete-case subset

	Full cohort ($n = 4711$)		**Complete cases ($n = 3336$)**	
los	Median [IQR]	5 [3–9]	Median [IQR]	5 [3–10]
	Mean	7.16	Mean	7.45
death	%	24.4%	%	23.5%
age	Median [IQR]	65 [54–76]	Median [IQR]	65 [55–76]
	Mean	63.4	Mean	64.3
oxygenation	Normal	55.5%	Normal	50.7%
	Mild	24.8%	Mild	27.2%
	Moderate/Severe	19.7%	Moderate/Severe	22.1%
map	Normal	72.0%	Normal	72.6%
	Low	9.8%	Low	9.9%
	High	18.2%	High	17.5%
creatinine	Normal	56.3%	Normal	55.4%
	Mild	20.7%	Mild	21.0%
	High	23.0%	High	23.6%
crp	Normal	30.2%	Normal	29.1%
	Mild	48.3%	Mild	49.0%
	High	21.5%	High	21.9%

Note. Length of stay (`los`) and age are summarized as median [IQR] and mean, whereas mortality is reported as a percentage. Other variables are shown as percentages within categories.

All the covariates listed in Table 3 were considered in the modelling of `los` and `death`. Their effects were specified through additive predictors, incorporating smooth functions for `age` to flexibly capture potential nonlinear associations. Categorical variables were represented using standard dummy (indicator) coding, with a separate parameter for each category. To maintain model parsimony and interpretability, interaction terms were not included; however, the approach readily accommodates them if specific interactions are of scientific interest or merit exploration.

For the binary response, various link functions were evaluated, while for the count outcome, multiple candidate distributions were explored. Model choice was guided by convergence diagnostics, information criteria and residual evaluation. Covariate selection then followed. Finally, the association between the outcomes was investigated using copulae, employing various families and additive predictor configurations to determine the dependence structure with the strongest empirical support.

The Akaike Information Criterion (AIC) (Akaike, 1998), here generically defined as $\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2edf$, was used throughout the model-building process. Residual analysis for the marginal models was performed using randomized quantile residuals, defined as $e_{ij} = \Phi^{-1}(u_{ij})$ for $i = 1, \ldots, n$ and $j = 1, 2$, where, for the binary outcome, u_{i1} was drawn from a uniform distribution on $[0, F_1(0; \hat{\mu}_{i1})]$ when $y_{i1} = 0$ and on $[F_1(0; \hat{\mu}_{i1}), 1]$ when $y_{i1} = 1$, whereas for the count outcome, u_{i2} was drawn from a uniform on $[F_2(y_{i2}; \hat{\mu}_{i2}, \hat{\sigma}_{i2}) - f_2(y_{i2}; \hat{\mu}_{i2}, \hat{\sigma}_{i2}),\ F_2(y_{i2}; \hat{\mu}_{i2}, \hat{\sigma}_{i2})]$ (Dunn & Smyth, 1996). Under correct model specification, $e_{ij} \overset{a}{\sim} \mathcal{N}(0, 1)$, assessed via normal Q–Q plots. It should be noted that, for binary outcomes, residuals are not very informative (e.g. Collett, 2002), and carrying out a sensitivity analysis using alternative link functions is generally preferable, although experience suggests that this choice typically does not substantially affect model fit.

Bivariate randomized quantile residuals (e.g. Hohberg et al., 2021) were used to perform an overall assessment of the copula regression models. They are defined here as $\mathbf{e}_i = (e_{i1}, e_{i2|1})^T$, where $e_{i2|1} = \Phi^{-1}(u_{i2|1})$ represents the conditional residual for the count outcome, with $u_{i2|1}$ drawn from a uniform distribution on

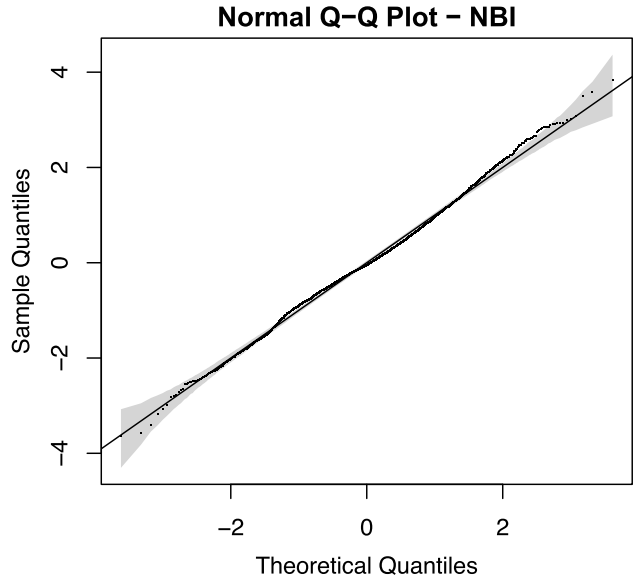

Figure 1. Normal Q–Q plot of randomized quantile residuals for the count outcome, derived from a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data. Shaded bands correspond to 95% pointwise reference intervals obtained from 1000 repeated samples from the standard normal distribution.

$$\begin{bmatrix} \frac{C(F_1(0;\hat{\mu}_{i1}),\, F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2})-f_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2});\hat{\theta}_i)}{F_1(0;\hat{\mu}_{i1})}, \\ \frac{C(F_1(0;\hat{\mu}_{i1}),\, F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2});\hat{\theta}_i)}{F_1(0;\hat{\mu}_{i1})} \end{bmatrix}, \quad \text{when } y_{i1}=0$$

and

$$\begin{bmatrix} \frac{F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2})-f_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2})-C(F_1(0;\hat{\mu}_{i1}),\, F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2})-f_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2});\hat{\theta}_i)}{F_1(1;\hat{\mu}_{i1})}, \\ \frac{F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2})-C(F_1(0;\hat{\mu}_{i1}),\, F_2(y_{i2};\hat{\mu}_{i2},\hat{\sigma}_{i2});\hat{\theta}_i)}{F_1(1;\hat{\mu}_{i1})} \end{bmatrix}, \quad \text{when } y_{i1}=1.$$

For a correctly specified joint model, $\mathbf{e}_i$ approximately follows a bivariate standard normal distribution with independent components. Since $e_{i1} \overset{a}{\sim} \mathcal{N}(0,1)$ and $e_{i2|1} \overset{a}{\sim} \mathcal{N}(0,1)$, it then follows that $\mathbf{e}_i^T\mathbf{e}_i \overset{a}{\sim} \chi^2(2)$, assessed via a Q–Q plot.

5.1.1 Marginal distributions

For the binary response, representing mortality and modelled using a Bernoulli distribution, the `probit`, `logit` and `cloglog` link functions were evaluated. Since all the candidate models yielded similar AIC values, the `logit` link was selected for its interpretability and familiarity among clinicians.

For the length of hospital stay, the distributions in Table 1 were assessed through normal Q–Q plots and convergence diagnostics. Overall, all the models except the Poisson exhibited satisfactory convergence and residual behaviour, with the Negative Binomial Type I showing the most well-behaved residuals.

The marginal models were then refined. In the μ_1 equation, the smooth term of `age` was replaced by a linear effect, as its *edf* equalled 1. In the μ_2 and σ_2 equations, the smooth terms of `age` displayed nonlinear patterns and were therefore retained. Backward selection based on *p*-values, using a 5% significance level, was then applied to remove covariates showing little statistical support. Figure 1 generally corroborates the chosen count distribution, despite some lack of fit in the upper tail.

5.1.2 Copula selection

The dependence between the chosen marginals was captured using the families listed in Table 2. A Gaussian copula was initially considered, with the correlation parameter expressed as a function of covariate effects. Since the estimated relationship was consistently positive across observations, the subsequent search was restricted to copulae permitting only positive dependence, including the Clayton and its 180-degree rotation, AMH, FGM, Frank, Joe and its 180-degree rotation, Plackett, and Student's *t* copula with degrees of freedom ranging from 3 to 6 (chosen as a representative range for the application considered). Given that the overall estimated correlation was weak to moderate, with a value of 0.29, the Gumbel and Galambos copulae were not considered, not for theoretical reasons, but because at this level of dependence they exhibit patterns very similar to those of the Clayton and Joe; including them would have added little new information while increasing the number of candidate models and the computational burden. The reduced set was deemed sufficient to capture the form of dependence in the case study; however, in applications demonstrating stronger association, the Gumbel and Galambos may offer additional insight. As for the Student's *t*, the lack of empirical support made it unnecessary to consider additional degrees of freedom. Once the copula most supported by the data was determined, the covariates for the dependence parameter were selected via backward elimination using *p*-values and a 5% significance level. Following a reviewer's suggestion, a simpler specification in which the dependence parameter was assumed constant, i.e. modelled using an intercept only, was also evaluated.

Copula selection was guided primarily by the AIC, which, as summarized by Czado (2019, Section 8.1), is widely recommended for establishing the copula family. Following the discussion in Czado (2019, Section 9.3), the Vuong test (Vuong, 1989) was also applied, which compares two competing models by computing the difference in their log-likelihoods, standardized by the corresponding standard error. Under the null hypothesis that two models are observationally equivalent, the resulting statistic follows a standard normal distribution.

Table 6 indicates that the Clayton (`C0`) with a varying dependence parameter provided the best fit according to the AIC. The alternative `J180` produced a virtually identical fit, which was expected given the level of association between the marginals and the fact that both `C0` and `J180` capture similar asymmetric dependence. This suggested that the substantive conclusions of the analysis were unlikely to be affected by the choice of asymmetric copula. The Vuong test, computed as the number of pairwise wins across all comparisons, broadly supported this conclusion. For the models with a constant dependence

Table 6. Comparison of copula models in terms of AIC and Vuong wins

Copula	AIC (covariate-dependent)	Vuong (covariate-dependent)	AIC (constant)	Vuong (constant)
C0	22,738.18	8	22,773.75	7
J180	22,738.87	8	22,777.75	6
F	22,740.59	8	22,784.63	5
N	22,740.40	7	22,767.80	10
PL	22,740.45	7	22,784.94	6
C180	22,770.54	3	22,789.09	3
T6	22,770.67	3	22,791.19	3
AMH	22759.38	2	22,784.27	4
T5	22,778.29	2	22,798.03	2
FGM	22762.10	1	22,783.69	5
J0	22,782.11	1	22,802.15	1
T4	22,790.15	1	22,808.88	1
T3	22,810.48	0	22,827.89	0

Note. Columns 2–3 report the AIC and number of Vuong wins for the models in which the dependence parameter is allowed to vary with covariate effects (covariate-dependent specification), whereas columns 4–5 report the same statistics for the simpler specification in which the dependence parameter is constant.

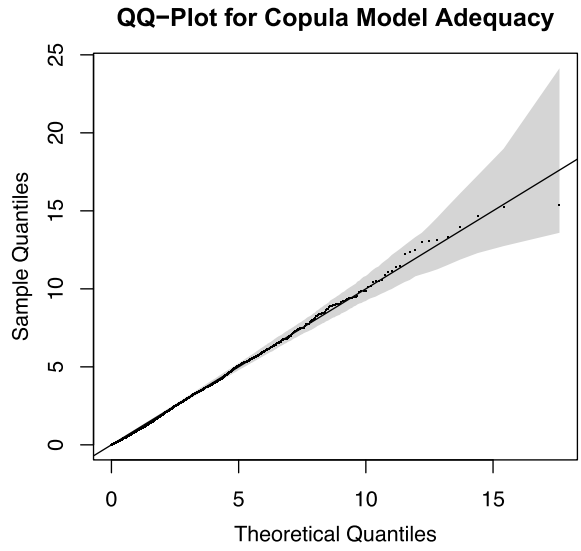

Figure 2. Q–Q plot of the sum of squared bivariate randomized quantile residuals against the $\chi^2(2)$ distribution, derived from a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data. Shaded bands correspond to 95% pointwise reference intervals obtained from 1000 repeated samples from the $\chi^2(2)$ distribution.

parameter, the situation differed: the `N` copula achieved the lowest AIC and highest number of Vuong wins, followed by `C0`. Nevertheless, the covariate-dependent specification was generally preferred, both based on the AIC and for substantive interpretation, as it allowed for a richer understanding of how dependence varies with covariate effects, providing more detailed insights into the structure of the joint distribution. The overall adequacy of the copula models was assessed using Q–Q plots of the sum of squared bivariate randomized quantile residuals, with Figure 2 supporting the selected model.

5.2 The final model

The chosen model is based on a Clayton copula combining a Bernoulli distribution with a `logit` link and a Negative Binomial Type I, where the model parameters are specified as $\Phi^{-1}(\mu_1) = \eta_{\mu_1}(\boldsymbol{x}_{\mu_1}; \boldsymbol{\beta}_{\mu_1})$, $\log(\mu_2) = \eta_{\mu_2}(\boldsymbol{x}_{\mu_2}; \boldsymbol{\beta}_{\mu_2})$, $\log(\sigma_2) = \eta_{\sigma_2}(\boldsymbol{x}_{\sigma_2}; \boldsymbol{\beta}_{\sigma_2})$ and $\log(\theta) = \eta_{\theta}(\boldsymbol{x}_{\theta}; \boldsymbol{\beta}_{\theta})$, with additive predictors given by

$$\begin{aligned}
\eta_{\mu_1}\left(\boldsymbol{x}_{\mu_1}; \boldsymbol{\beta}_{\mu_1}\right) &= \beta_{0\mu_1} + \beta_{1\mu_1}\texttt{pure.cns}_{\text{Yes}} + \beta_{2\mu_1}\texttt{age} + \beta_{3\mu_1}\texttt{oxygenation}_{\text{Mild}} + \beta_{4\mu_1}\texttt{oxygenation}_{\text{Mod/Sev}} \\
&\quad + \beta_{5\mu_1}\texttt{temperature}_{\text{Hypothermia}} + \beta_{6\mu_1}\texttt{temperature}_{\text{Fever}} + \beta_{7\mu_1}\texttt{map}_{\text{Low}} + \beta_{8\mu_1}\texttt{map}_{\text{High}} \\
&\quad + \beta_{9\mu_1}\texttt{creatinine}_{\text{Mild}} + \beta_{10\mu_1}\texttt{creatinine}_{\text{High}} + \beta_{11\mu_1}\texttt{ast}_{\text{High}} + \beta_{12\mu_1}\texttt{lymphocytes}_{\text{Low}} \\
&\quad + \beta_{13\mu_1}\texttt{lymphocytes}_{\text{High}} + \beta_{14\mu_1}\texttt{crp}_{\text{Mild}} + \beta_{15\mu_1}\texttt{crp}_{\text{High}} + \beta_{16\mu_1}\texttt{troponin}_{\text{Elevated}}, \\
\eta_{\mu_2}\left(\boldsymbol{x}_{\mu_2}; \boldsymbol{\beta}_{\mu_2}\right) &= \beta_{0\mu_2} + \beta_{1\mu_2}\texttt{pure.cns}_{\text{Yes}} + s_{1\mu_2}(\text{age}) + \beta_{2\mu_2}\texttt{oxygenation}_{\text{Mild}} + \beta_{3\mu_2}\texttt{oxygenation}_{\text{Mod/Sev}} \\
&\quad + \beta_{4\mu_2}\texttt{temperature}_{\text{Hypothermia}} + \beta_{5\mu_2}\texttt{temperature}_{\text{Fever}} + \beta_{6\mu_2}\texttt{map}_{\text{Low}} + \beta_{7\mu_2}\texttt{map}_{\text{High}} \\
&\quad + \beta_{8\mu_2}\texttt{creatinine}_{\text{Mild}} + \beta_{9\mu_2}\texttt{creatinine}_{\text{High}} + \beta_{10\mu_2}\texttt{sodium}_{\text{Hyponatremia}} \\
&\quad + \beta_{11\mu_2}\texttt{sodium}_{\text{Hypernatremia}} + \beta_{12\mu_2}\texttt{lymphocytes}_{\text{Low}} + \beta_{13\mu_2}\texttt{lymphocytes}_{\text{High}} \\
&\quad + \beta_{14\mu_2}\texttt{crp}_{\text{Mild}} + \beta_{15\mu_2}\texttt{crp}_{\text{High}}, \\
\eta_{\sigma_2}\left(\boldsymbol{x}_{\sigma_2}; \boldsymbol{\beta}_{\sigma_2}\right) &= \beta_{0\sigma_2} + s_{1\sigma_2}(\text{age}) + \beta_{1\sigma_2}\texttt{oxygenation}_{\text{Mild}} + \beta_{2\sigma_2}\texttt{oxygenation}_{\text{Mod/Sev}} \\
&\quad + \beta_{3\sigma_2}\texttt{map}_{\text{Low}} + \beta_{4\sigma_2}\texttt{map}_{\text{High}} + \beta_{5\sigma_2}\texttt{sodium}_{\text{Hyponatremia}} + \beta_{6\sigma_2}\texttt{sodium}_{\text{Hypernatremia}} \\
&\quad + \beta_{7\sigma_2}\texttt{lymphocytes}_{\text{Low}} + \beta_{8\sigma_2}\texttt{lymphocytes}_{\text{High}}
\end{aligned}$$

and

$$\begin{aligned}\eta_\theta(\boldsymbol{x}_\theta;\boldsymbol{\beta}_\theta) = {} & \beta_{0\theta} + \beta_{1\theta}\texttt{age} + \beta_{2\theta}\texttt{oxygenation}_{\text{Mild}} + \beta_{3\theta}\texttt{oxygenation}_{\text{Mod/Sev}} \\ & + \beta_{4\theta}\texttt{map}_{\text{Low}} + \beta_{5\theta}\texttt{map}_{\text{High}} + \beta_{6\theta}\texttt{creatinine}_{\text{Mild}} + \beta_{7\theta}\texttt{creatinine}_{\text{High}} \\ & + \beta_{8\theta}\texttt{crp}_{\text{Mild}} + \beta_{9\theta}\texttt{crp}_{\text{High}}.\end{aligned}$$

Model convergence was satisfactory, with the maximum absolute gradient effectively equal to zero and a positive definite observed information matrix.

5.2.1 Model fitting in R

The copula modelling framework is available in the R package GJRM (Marra & Radice, 2026), which facilitates model estimation and inference, and provides easily interpretable numerical and graphical summaries. The model can be implemented using the following code

```
library(GJRM)
eqmu1 <- death  ~  pure.cns + age + oxygenation + temperature + map +
                   creatinine + ast + lymphocytes + crp + troponin
eqmu2 <- los    ~  pure.cns + s(age) + oxygenation + temperature + map +
                   creatinine + sodium + lymphocytes + crp
eqsi2 <-        ~  s(age) + oxygenation + map + sodium + lymphocytes
eqthe <-       ~  age + oxygenation + map + creatinine + crp

outC0 <- gjrm(list(eqmu1, eqmu2, eqsi2, eqthe), data = hosp,
              margins = c("logit", "NBI"), copula = "C0", model = "B")
```

where the various equations have the obvious interpretations, the `data` argument indicates the dataset used, `margins` defines the marginal distributions, `copula` specifies the copula function adopted to model the dependence between the responses and `model` = "B" indicates that a bivariate model is being fitted. Post-estimation functions such as `conv.check()`, `copula.prob()`, `vuong.test()`, `summary()` and `plot()` can be employed to check for convergence, perform model comparisons and extract numerical and visual summaries.

5.2.2 Covariate effects on μ_1

Several clinical and laboratory variables as well as `age` had an impact on in-hospital mortality. The related estimated coefficients, standard errors and *p*-values are presented in Table 7. Specifically, `age` was a strong predictor, with each additional year increasing the odds of death (OR) by approximately 4%, consistent with the fact that older patients face higher COVID-19 mortality. Patients with neurological involvement (`pure.cns`) also had a higher risk of death (OR = 1.45), reflecting the impact of central nervous system complications such as stroke or encephalopathy. Respiratory status was a key determinant of the outcome: moderate to severe hypoxemia nearly doubled the OR, whereas mild hypoxemia did not significantly alter the risk, aligning with clinical expectations that severe hypoxemia is linked with extensive lung injury. Hemodynamic instability emerged as a strong predictor: `map` <65 mmHg was associated with more than an eleven-fold higher OR, while elevated `map` was not linked to an increased risk. Renal dysfunction was also associated with mortality. Both mild and severe `creatinine` elevations increased the odds of death (OR = 1.73 and OR = 1.86, respectively), in line with medical evidence that acute kidney injury signals severe disease. Elevated `ast` conferred a modest increase in the risk (OR = 1.25), suggesting hepatic involvement or systemic stress. Inflammatory markers exhibited a dose-response relationship with mortality. Mildly elevated `crp` increased the odds of death 1.75 times, whereas `crp` >20 mg/L nearly tripled the odds (OR = 2.76), consistent with the role of systemic inflammation in adverse outcomes. Lymphopenia ($<1\times10^3/\mu$L) also influenced mortality (OR = 1.30), whereas lymphocytosis ($>3.5\times10^3/\mu$L) was rare and not associated

Table 7. Estimated coefficients for μ_1, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data

Death model - μ_1 parameter			
Parameter	**Estimate**	**Std. error**	***p*-value**
(Intercept)	-5.794	0.294	<0.001
pure.cnsYes	0.375	0.147	0.011
age	0.042	0.004	<0.001
oxygenationModerate/Severe hypoxemia (<=89%)	0.686	0.120	<0.001
oxygenationMild hypoxemia (90–94%)	0.034	0.121	0.776
temperatureHypothermia (<36°C)	0.057	0.304	0.852
temperatureFever (>=38°C)	0.221	0.117	0.059
mapLow (<65 mmHg)	2.438	0.151	<0.001
mapHigh (>100 mmHg)	0.066	0.130	0.611
creatinineMildly elevated (1.3–2.0 mg/dL)	0.548	0.119	<0.001
creatinineHigh (>2.0 mg/dL)	0.618	0.132	<0.001
astHigh (>40 U/L)	0.222	0.098	0.024
lymphocytesLow (<1 x10^3/μL)	0.265	0.098	0.007
lymphocytesHigh (>3.5 x10^3/μL)	0.077	0.426	0.857
crpMildly elevated (5–20 mg/L)	0.558	0.135	<0.001
crpHigh (>20 mg/L)	1.017	0.152	<0.001
troponinElevated (>0.04 ng/mL)	0.267	0.130	0.039

with the probability of death. Finally, myocardial injury, indicated by elevated `troponin`, was linked with a 31% increase in the OR, corroborating the prognostic relevance of cardiac involvement in hospitalized patients.

5.2.3 Covariate effects on μ_2 and σ_2

A number of patient characteristics were associated with the expected duration of hospitalization, as captured by the μ_2 parameter (see Table 8). Patients with neurological complications had an expected 29% longer stay, consistent with the need for more prolonged care in this subgroup. Respiratory status also influenced `los`: moderate to severe hypoxemia resulted in a 20% longer duration, whereas mild hypoxemia increased the expected length by about 8%. Abnormalities in `temperature` had more modest effects. Fever corresponded to an estimated 10% longer expected stay, while hypothermia tended toward shorter admissions (-18%). Hemodynamic extremes also contributed: hypotension increased the expected stay by approximately 12%, whereas elevated `map` was associated with a non-significant trend toward shorter hospitalization. Among laboratory markers, severe renal dysfunction was linked with an 18% longer duration, whereas mild elevations did not significantly alter the outcome. Electrolyte disturbances had smaller effects: hyponatremia modestly increased the expected stay by 7%, while hypernatremia had no significant impact. Inflammatory and immune markers were important predictors: both mildly elevated and high `crp` were linked with roughly 22% longer hospitalizations, reflecting the contribution of systemic inflammation to prolonged care. Lymphopenia was associated with an 8% increase in the expected stay, whereas the effect of lymphocytosis was rare and not significant.

The variability of `los`, captured by the σ_2 parameter, was influenced by a few covariates, in different directions (see Table 9). Mild hypoxemia led to reduced variability, suggesting that patients with modest oxygen deficits had more predictable hospital courses. A similar pattern was observed for high `map`. Conversely, lymphopenia increased variability, and extreme lymphocytosis was associated with markedly greater uncertainty, possibly reflecting heterogeneous responses in this subset of patients. Other factors, including hyponatremia and severe hypoxemia, did not significantly alter this parameter.

Table 8. Estimated coefficients for μ_2, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data

Length of stay model - μ_2 parameter			
Parameter	**Estimate**	**Std. error**	***p*-value**
(Intercept)	1.655	0.038	<0.001
pure.cnsYes	0.258	0.049	<0.001
oxygenationModerate/Severe hypoxemia (<=89%)	0.179	0.041	<0.001
oxygenationMild hypoxemia (90–94%)	0.076	0.035	0.032
temperatureHypothermia (<36°C)	-0.199	0.105	0.058
temperatureFever (> = 38°C)	0.097	0.038	0.010
mapLow (<65 mmHg)	0.111	0.050	0.027
mapHigh (>100 mmHg)	-0.071	0.039	0.065
creatinineMildly elevated (1.3–2.0 mg/dL)	-0.007	0.039	0.847
creatinineHigh (>2.0 mg/dL)	0.163	0.038	<0.001
sodiumHyponatremia (<135 mmol/L)	0.068	0.034	0.045
sodiumHypernatremia (>145 mmol/L)	0.020	0.053	0.703
lymphocytesLow (<1 x10$^3/\mu$L)	0.073	0.031	0.018
lymphocytesHigh (>3.5 x10$^3/\mu$L)	−0.075	0.181	0.679
crpMildly elevated (5–20 mg/L)	0.198	0.037	<0.001
crpHigh (>20 mg/L)	0.197	0.046	<0.001

Table 9. Estimated coefficients for σ_2, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data

Length of stay model - σ_2 parameter			
Parameter	**Estimate**	**Std. Error**	***p*-value**
(Intercept)	−0.491	0.060	<0.001
oxygenationModerate/Severe hypoxemia (< = 89%)	0.007	0.079	0.926
oxygenationMild hypoxemia (90–94%)	−0.208	0.074	0.005
mapLow (<65 mmHg)	−0.131	0.103	0.203
mapHigh (>100 mmHg)	−0.241	0.087	0.005
sodiumHyponatremia (<135 mmol/L)	0.027	0.069	0.694
sodiumHypernatremia (>145 mmol/L)	0.200	0.104	0.054
lymphocytesLow (<1 x10$^3/\mu$L)	0.158	0.063	0.012
lymphocytesHigh (>3.5 x10$^3/\mu$L)	0.957	0.232	<0.001

Figure 3 reveals nonlinear relationships with length of stay and its variability. For the expected duration, the effect of `age` followed an inverted-U shape: the expected stay increased with `age` up to approximately 50 years, after which it gradually declined. This pattern is plausible for COVID-19 hospitalizations, as middle-aged patients may experience a combination of comorbidities and disease severity that prolongs stay, whereas very young patients generally recover quickly and older patients may have higher early mortality, effectively shortening hospitalization. Regarding the σ_2 parameter, `age` above 50 years was associated with a modest decrease in the variability of the outcome, possibly suggesting that, among older patients, the course of illness and its management were more homogeneous. Overall, these nonlinear effects are consistent with clinical and epidemiological knowledge, demonstrating that `age` influences both the expected duration and the heterogeneity of hospital stay in a complex but interpretable manner.

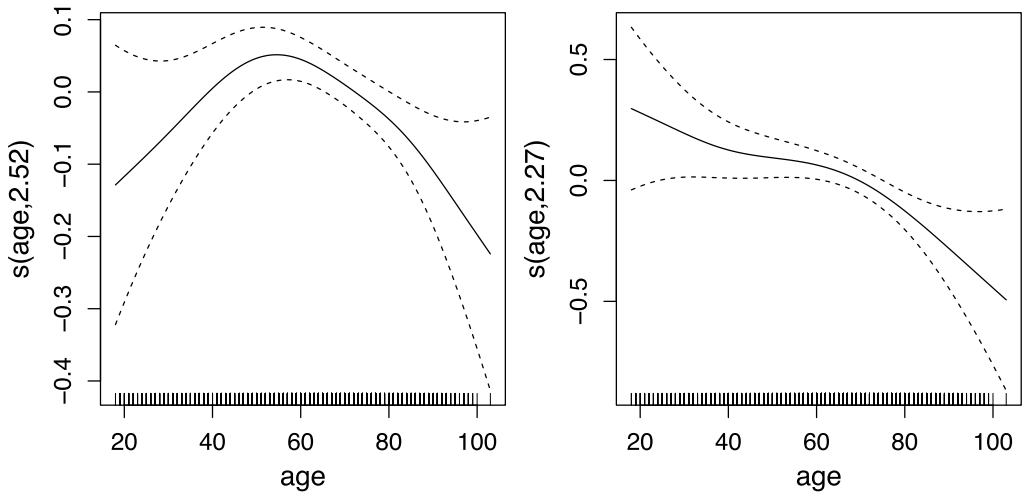

Figure 3. Estimated smooth effects (with associated 95% intervals) of `age` on the scale of the additive predictors of μ_2 and σ_2, respectively, derived from a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data.

Table 10. Estimated coefficients for the copula parameter, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data

Copula dependence parameter θ			
Parameter	**Estimate**	**Std. Error**	***p*-value**
(Intercept)	3.229	0.586	<0.001
age	-0.039	0.008	<0.001
oxygenationModerate/Severe hypoxemia (<=89%)	-1.277	0.360	<0.001
oxygenationMild hypoxemia (90-94%)	-0.640	0.305	0.036
mapLow (<65 mmHg)	0.532	0.294	0.070
mapHigh (>100 mmHg)	-1.129	0.538	0.036
creatinineMildly elevated (1.3-2.0 mg/dL)	-0.213	0.295	0.470
creatinineHigh (>2.0 mg/dL)	-0.926	0.328	0.005
crpMildly elevated (5-20 mg/L)	-0.347	0.282	0.218
crpHigh (>20 mg/L)	-0.705	0.352	0.045

5.2.4 Covariate effects on θ

The relationship between the responses was modelled using a Clayton copula with a covariate-dependent θ. The overall parameter was $\hat{\theta} = 1.13$, with a 95% interval of (0.556, 2.33). This indicates that the probability of discharge and length of stay were positively associated, with stronger dependence between lower likelihood of discharge and shorter lengths of stay. Equivalently, if the focus is on the probability of death and length of stay, the copula structure corresponds to a 90-degree rotated Clayton, implying a negative dependence between them, with stronger association between higher mortality risk and shorter hospitalizations and a weaker link between lower likelihood of death and longer stays.

Several covariates influenced the copula parameter (see Table 10). In particular, the magnitude of θ decreased progressively with `age`, indicating that the link between the probability of discharge and hospitalization duration diminished in older individuals. Similarly, moderate to severe hypoxemia

markedly reduced dependence, and mild hypoxemia had a smaller but significant effect. Severe renal dysfunction and markedly elevated `crp` also decreased θ, indicating weaker coupling between hospital stay and the likelihood of discharge in these subgroups. In contrast, hypotension exhibited a positive, although marginally significant, impact.

5.3 Joint probabilities

Figures 4 and 5 illustrate the joint probabilities of in-hospital `death` and `los`, as well as discharge and `los`, for two contrasting patient profiles evaluated at ages 18, 30, 60 and 90 years. These profiles were constructed to represent clinically interpretable and realistic subject scenarios within the observed cohort.

The low-risk profile represents a clinically stable patient without central nervous system involvement and with physiological and biochemical measurements within normal clinical ranges. Covariates were fixed at reference values, including oxygen saturation ($\geq$ 95%), temperature (36–37.9°C), mean arterial pressure (65–100 mmHg), creatinine ($\leq$ 1.2 mg/dL), AST ($\leq$ 40 U/L), lymphocyte count (1–3.5 $\times 10^3/\mu$L), CRP (<5 mg/L), troponin ($\leq$ 0.04 ng/mL) and sodium (135–145 mmol/L).

The high-risk profile corresponds to a clinically severe but realistic COVID-19 presentation, characterized by the presence of at least three adverse features commonly observed among critically ill patients in the cohort. Specifically, this profile includes moderate to severe hypoxemia (<90%), fever ($\geq$ 38°C), elevated creatinine (>2.0 mg/dL) and markedly elevated CRP (>20 mg/L), while other covariates are set to typical values. Approximately 6% of individuals in the complete-case sample exhibited this combination of three or more risk features, supporting the clinical plausibility of the profile.

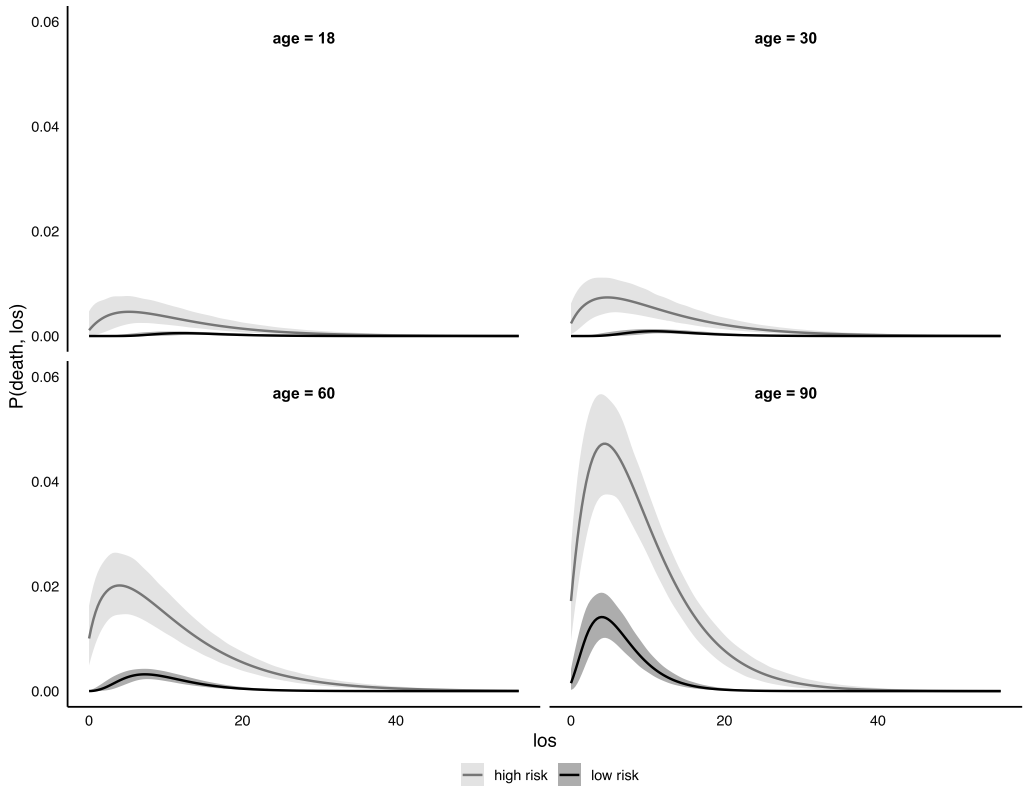

Figure 4. Joint probability estimates of `death` and `los` by `age` and patient profile, with 95% pointwise intervals, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data.

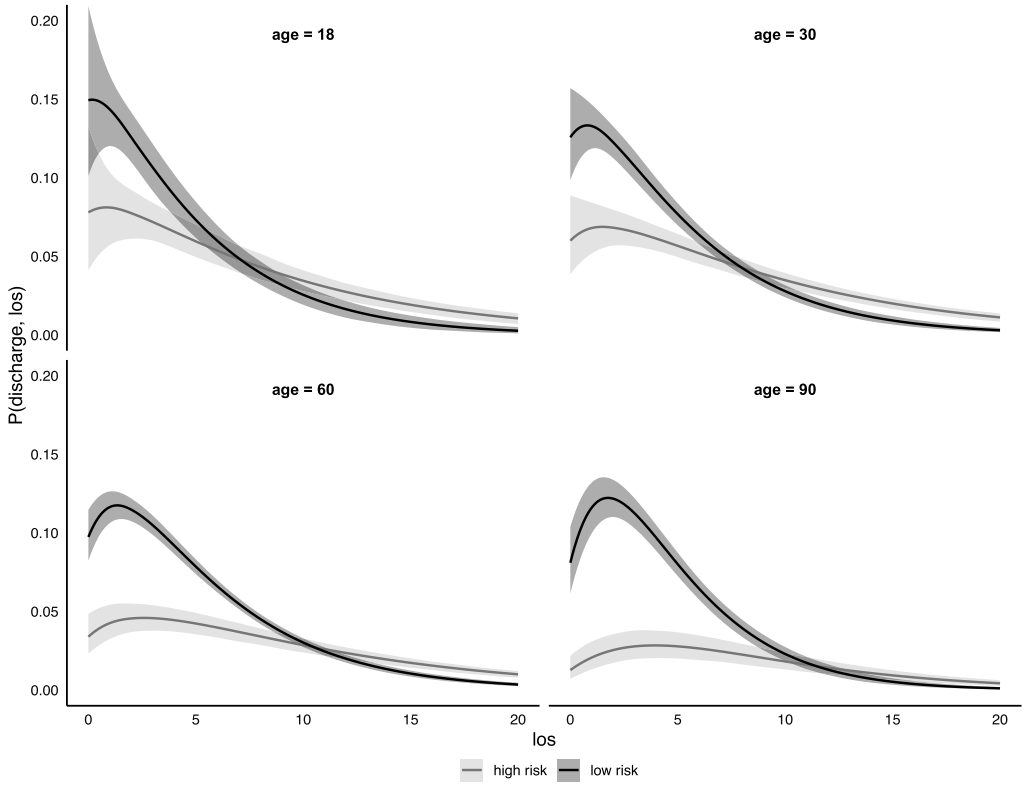

Figure 5. Joint probability estimates of discharge and `los` by `age` and patient profile, with 95% pointwise intervals, based on a Clayton copula additive distributional regression model with Bernoulli (`logit` link) and `NBI` margins fitted to COVID-19 hospital data.

For low-risk subjects, the joint probability of `death` and `los` remained consistently low across varying lengths of stay, indicating that fatal outcomes and extended admissions rarely coincided within this group. A modest increase occurred from approximately age 60, reaching a peak at 90 for a `los` of around four days: 0.014 with 95% interval (0.010, 0.019). Conversely, high-risk patients displayed substantially higher joint probabilities overall, with values increasing progressively with age. Among this cohort, death and length of stay were more likely to co-occur at shorter durations, with the highest estimated joint probability observed for individuals aged 90 and a `los` of approximately five days: 0.047 (0.037, 0.057).

With respect to discharge, joint probabilities for high-risk patients were highest at younger ages and very short hospitalization durations. For example, at age 18 and `los` = 2 days, the estimated joint probability was 0.081 (0.058, 0.105), whereas for an otherwise comparable subject aged 90 years it decreased to 0.025 (0.017, 0.037). Younger low-risk individuals exhibited substantially higher joint probabilities of discharge, particularly for minimal lengths of stay. At age 18 and `los` = 2, the joint probability for a low-risk patient, estimated at 0.130 (0.109,0.141), was notably higher than that of the corresponding high-risk one. As `los` increased, joint discharge probabilities declined for both profiles, and the differences between low- and high-risk subjects became progressively smaller.

5.3.1 Practical insights

The joint probability estimates highlight clear differences in how the outcomes co-occurred across patient profiles and ages. Among high-risk individuals, particularly at advanced ages, the joint probability of in-hospital death and length of stay was concentrated at short to moderate hospitalization

durations, peaking at around five days for patients aged 90. This indicates that fatal outcomes in this group were mainly observed early during admission, emphasizing the importance of intensifying monitoring and timely intervention soon after hospital entry. In contrast, low-risk subjects exhibited consistently low joint probabilities of death across all durations, with only modest increases at older ages. Extended hospitalizations in this group rarely coincided with fatal outcomes, suggesting that prolonged stays were more likely driven by nonlethal complications or nonclinical factors. These patterns suggest that intensive resources might have been prioritized for patients whose risk was highest early in admission.

Joint probabilities of discharge and hospitalization duration were highest at very short stays, particularly among younger low-risk patients. For example, at age 18 with a two-day stay, low-risk subjects were substantially more likely to be discharged than comparable high-risk individuals. As length of stay increased, discharge probabilities declined for both profiles, and the differences between groups narrowed. Operationally, these findings indicate that early discharge pathways could have been implemented for low-risk patients, with front-loaded allocation of critical care resources for older, high-risk individuals.

6 Conclusions and future directions

This paper employed a copula additive distributional regression framework to jointly model in-hospital mortality and length of stay, capturing the complex dependencies between these key healthcare metrics. By allowing all the parameters of the bivariate distribution to be modelled through additive predictors, the approach provided a nuanced understanding of how patient characteristics and clinical measurements shaped the joint behaviour of mortality (or discharge) and hospitalization duration among patients admitted with COVID-19 during the first surge of the pandemic in New York City. All model fitting and subsequent analyses were conducted using the `GJRM` package in `R`. Despite the underlying complexity of the methodology, the package streamlines estimation and interpretation, enabling applied researchers and healthcare professionals to derive interpretable and practically relevant results.

From a methodological perspective, the framework can be extended to jointly examine additional outcomes such as ICU admission, readmission and other clinical endpoints. Incorporating ICU admissions would offer deeper insights into disease severity and critical resource utilization, while modelling readmissions would improve the understanding of post-discharge risk and continuity of care.

Beyond the hospital setting, the proposed framework can be employed in several operational contexts where multiple outcomes are connected. In chronic disease management, the model may be used to characterize the dependencies among treatment adherence, disease progression and healthcare utilization, informing long-term care strategies. In public health, it can quantify the interconnected dynamics of infection rates, vaccination uptake and hospitalization, supporting more effective intervention planning and resource allocation.

Finally, future work will explore the development of predictive validation tools, such as out-of-sample scoring procedures, to assess the performance and reliability of copula regression models in evolving clinical populations. In addition, further methodological research could investigate multiple-comparison correction procedures specifically tailored to penalized copula additive regression models, where classical approaches may not be directly applicable.

Acknowledgments

The authors thank the editors and the two anonymous reviewers for their insightful and constructive comments, which helped strengthen the clarity and overall quality of the paper.

Conflicts of interest

None declared.

Funding

This research received no specific grant from any funding agency.

Data availability

The dataset used in this study is publicly available at: https://figshare.com/s/79827c396af7df42b3d7?file=24020852.

References

Akaike H. (1998). *Information theory and an extension of the maximum likelihood principle*. In *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.

Altschul D. J., Unda S. R., Benton J., Cox M., Dardick J., De La Garza Ramos R., Greenfield J. P., Magge S. N., & Schwartz T. H. (2020). A novel severity score to predict inpatient mortality in COVID-19 patients. *Scientific Reports*, *10*, 16726. https://doi.org/10.1038/s41598-020-73962-9

Chen J., Qi T. D., Vu J., & Wen Y. (2023). A deep learning approach for inpatient length of stay and mortality prediction. *Journal of Biomedical Informatics*, *147*, 104526. https://doi.org/10.1016/j.jbi.2023.104526

Collett D. (2002). *Modelling binary data* (2nd ed.). Chapman & Hall/CRC.

Czado C. (2019). *Analyzing dependent data with vine copulas: A practical guide with R* (Vol. Lecture Notes in Statistics. *222*). Springer.

Dunn P. K., & Smyth G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *5*(3), 236–244. https://doi.org/10.1080/10618600.1996.10474708

Griffiths P., Maruotti A., Saucedo A. R., Redfern O. C., Ball J. E., Briggs J., Dall'Ora C., Schmidt P. E., & Smith G. B. (2019). Nurse staffing, nursing assistants and hospital mortality: Retrospective longitudinal cohort study. *BMJ Quality & Safety*, *28*(8), 609–617. https://doi.org/10.1136/bmjqs-2018-008043

Han T. S., Murray P., Robin J., Wilkinson P., Fluck D., & Fry C. H. (2022). Evaluation of the association of length of stay in hospital and outcomes. *International Journal for Quality in Health Care*, *34*(2), 1–9. https://doi.org/10.1093/intqhc/mzab160

Hohberg M., Donat F., Marra G., & Kneib T. (2021). Beyond unidimensional poverty analysis using distributional copula models for mixed ordered-continuous outcomes. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *70*, 1365–1390. https://doi.org/10.1111/rssc.12517

Hunt-O'Connor C., Moore Z., Patton D., Nugent L., Avsar P., & O'Connor T. (2021). The effect of discharge planning on length of stay and readmission rates of older adults in acute hospitals: A systematic review and meta-analysis of systematic reviews. *Journal of Nursing Management*, *29*(8), 2697–2706. https://doi.org/10.1111/jonm.13409

Institute for Healthcare Improvement (2024). Achieving hospital-wide patient flow: The right care, in the right place, at the right Time. https://qi.elft.nhs.uk/wp-content/uploads/2025/09/IHIAchievingHospitalWidePatientFlowWhitePaper.pdf

Joe H. (2014). *Dependence modeling with copulas*. Chapman & Hall/CRC.

Kim A., & Covey C. (2022). Benefits of individualized discharge plans for hospitalized patients. *American Family Physician*, *106*(5), 500–501. https://www.proquest.com/scholarly-journals/benefits-individualized-discharge-plans/docview/2736160663/se-2?accountid=14511

Lee J., & Park J. (2025). Relationships among comorbidities, disease severity, and hospitalization duration in the united states using the healthcare cost and utilization project (hcup) database. *Journal of Clinical Medicine*, *14*(3), 680. https://doi.org/10.3390/jcm14030680

Lingsma H. F., Bottle A., Middleton S., Kievit J., Steyerberg E. W., & van de Mheen P. J. M. (2018). Evaluation of hospital outcomes: The relation between length-of-stay, readmission, and mortality

in a large international administrative database. *BMC Health Services Research*, *18*(1), 1–10. https://doi.org/10.1186/s12913-018-2916-1

Marra G., & Radice R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, *115*(530), 886–895. https://doi.org/10.1080/01621459.2019.1593178

Marra G., & Radice R. (2025). *Copula additive distributional regression using R* (1st ed.). Chapman and Hall/CRC.

Marra G., & Radice R. (2026). Gjrm: Generalized joint regression modeling. R package version 0.2-6.9.

Marra G., Radice R., & Zimmer D. M. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *69*(4), 953–971. https://doi.org/10.1111/rssc.12419

Marra G., & Wood S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*(1), 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x

Nelsen R. B. (2006). *An introduction to copulas*. Springer Series in Statistics.

Nocedal J., & Wright S. J. (2006). *Numerical optimization*. Springer-Verlag.

Organisation for Economic Co-operation and Development (2024). Length of hospital stay. https://www.oecd.org/en/data/indicators/length-of-hospital-stay.html.

Rochon P. A., Katz J. N., Morrow L. A., McGlinchey-Berroth R., Ahlquist M. M., Sarkarati M., & Minaker K. L. (1996). Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability: A prospective comparison of three comorbidity indices. *Medical Care*, *34*(11), 1093–1101. https://doi.org/10.1097/00005650-199611000-00004

Stasinopoulos M. D., Rigby R. A., Heller G. Z., Voudouris V., & Bastiani F. D. (2017). *Flexible regression and smoothing using GAMLSS in R*. Chapman & Hall/CRC.

Stone K., Zwiggelaar R., Jones P., & Parthaláin N. M. (2022). A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, *1*(4), e0000017. https://doi.org/10.1371/journal.pdig.0000017

Trivedi P.K., & Zimmer D.M. (2007). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, *1*(1), 1–111. https://doi.org/10.1561/0800000005

Vuong Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Wilder A. V., Cox B., Ridder D. D., Tambeur W., Boer G. V., Brouwers J., Claessens F., Bruyneel L., & Vanhaecht K. (2022). A comprehensive analysis of temporal trends of between-hospital variation in mortality, readmission and length of stay using logistic regression. *Healthcare Analytics*, *2*, 100123. https://doi.org/10.1016/j.health.2022.100123

Wood S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC.

Wood S. N., Pya N., & Säfken B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

World Health Organization (2024). Patient safety: global action on patient safety. https://www.who.int/news-room/fact-sheets/detail/patient-safety.

Wu Y., Fung H., Shum H.-M., Zhao S., Wong E. L.-Y., Chong K.-C., Hung C.-T., & Yeoh E.-K. (2022). Evaluation of length of stay, care volume, in-hospital mortality, and emergency readmission rate associated with use of diagnosis-related groups for internal resource allocation in public hospitals in hong kong. *JAMA Network Open*, *5*(2), e2145685. https://doi.org/10.1001/jamanetworkopen.2021.45685

Yang L., Frees E. W., & Zhang Z. (2020). Nonparametric estimation of copula regression models with discrete outcomes. *Journal of the American Statistical Association*, *115*(530), 707–720. https://doi.org/10.1080/01621459.2018.1546586

Yen H.-Y., Chi M.-J., & Huang H.-Y. (2022). Effects of discharge planning services and unplanned readmissions on post-hospital mortality in older patients: A time-varying survival analysis. *International Journal of Nursing Studies*, *128*, 104175. https://doi.org/10.1016/j.ijnurstu.2022.104175