

A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant

Philippe Lambert^{1,*,†} and François Vandenhende^{1,2}

¹*Université Catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20,
B-1348 Louvain-la-Neuve, Belgium*

²*Statistics & Information Sciences, Eli Lilly & Company, B-1348 Mont-Saint-Guibert, Belgium*

SUMMARY

A new model for multivariate non-normal longitudinal data is proposed. In a first step, each longitudinal series of data corresponding to a given response is modelled separately using a copula to relate the marginal distributions of the response at each time of observation. In a second step, at each observation time, the conditional (on the past) distributions of each response are related using another copula describing the relationship between the corresponding variables. Note that there is no need to consider the same family of distributions for these response variables. The technique is illustrated in a dose titration safety study on a new antidepressant. The haemodynamic effect on diastolic blood pressure, systolic blood pressure and heart rate is studied. These three responses are measured repeatedly over time on ten healthy volunteers during the dose escalation. The available covariates are sex and the concentration of drug in the plasma at time of measurement. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: longitudinal; multivariate; copula; data of mixed types

1. INTRODUCTION

Modelling multivariate longitudinal data is an important challenge in biostatistics. In the literature, most of the time the considered models assume that the original data or some transformation of them are multivariate normal with a variance-covariance matrix structured [1] to describe serial dependence, heterogeneity and dependence between response variables (see Section 4.1). Unfortunately, normality is certainly not a rule in practice. When the responses are not normal or/and when their marginal distributions are not in the same family, alternatives to the multivariate normal must be found. In this paper, we shall focus on longitudinal data of mixed types.

* Correspondence to: Philippe Lambert, Université Catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

† E-mail: lambert@stat.ucl.ac.be

Very few tools exist in the literature to model such data. Zeger and Liang [2] consider, at each time point t , a model for the mean of each outcome conditional on its history and on the values of the other outcomes at time t . The conditional variance of each outcome is assumed proportional to a fixed function of the conditional mean. An extension of the generalized estimating equations (GEE) [3] method is used to estimate parameters.

Jørgensen *et al.* [4] propose state-space models based on the class of Tweedie exponential dispersion models. They can deal with a few discrete distributions, namely the Poisson and the compound Poisson, and several continuous distributions, among which are the normal, the gamma and the inverse Gauss. Their vectors of observations observed repeatedly at equally spaced times are assumed conditionally independent given a univariate latent process.

More tools can be found in the literature to deal with data of mixed types when the vectors of observations are only observed once on independent or clustered units.

A joint model for mixed continuous and categorical data was proposed by Olkin and Tate [5]. It was extended to deal with missing values in reference [6]. Conditionally on the categorical responses combination (with a marginal multinomial distribution), they assume that the continuous variables are multivariate normal. The EM algorithm [7] is used to obtain the maximum likelihood estimators (MLEs) in the presence of missing values. Liu and Rubin [8] generalized the approach by considering a multivariate t distribution instead of the more constraining multivariate normal and by allowing different covariance matrices across cells.

Many models for mixed type data can be found in the literature devoted to the analysis of toxicity studies. These usually involve clustered responses of mixed types [9]. One example is in reproduction toxicology. The examination of the foetuses of an exposed pregnant mice provides data on the presence of foetus malformation, foetal death and foetal weight. These response variables of mixed types can then be related to dam size and exposure in a single model.

Fitzmaurice and Laird [10] proposed modelling such data using marginal regression models for each of the responses while treating the intracluster correlation as a nuisance. GEEs are derived to obtain consistent parameter estimates while accounting for intracluster correlation. The parameters from the marginal regression models are robust to dependence misspecification.

Another possibility would be to express the joint distribution of two (say) responses as the product of one marginal distribution and of the conditional distribution of the other response given the first one. The choice of the conditioning response is of course arbitrary. See reference [11] for a model in which a binary response is modelled conditionally on a continuous normal response and reference [12] for a discussion of the above factorization strategy.

In a similar problem, Catalano and Ryan [13] proposed modelling the discrete outcome using a latent variable. They considered a bivariate normal distribution for the latent variable and the continuous response.

The same type of approach with a single conditioning latent variable was used by Sammel *et al.* [14] to relate mixed type outcomes modelled using generalized linear models. It was used to analyse the effect of an anticonvulsant on the development of birth defects.

Regan and Catalano [15] showed how to build a joint model for clustered binary and continuous outcomes with an application to developmental toxicology. They use an extension of the correlated probit model [16] to obtain a joint distribution for the binary and the continuous responses. A joint multivariate normal distribution is considered for the corresponding latent variables and each outcome is analysed with a marginal dose-response

model. The covariance matrix takes into account the correlation between outcomes and the correlation due to clustering. This is an important improvement of references [13] and [10] as model estimates of the correlation between responses and the evolution of these correlations with dose are available. A marginal approach was preferred to conditional models with the arguments that the biological mechanism relating the chemical exposition to developmental problems is not well understood, and that conditional models are difficult to use to quantify risk or to study the marginal effect of dose on an isolated response.

In this work, we propose to show how copulas can be used to model multivariate non-normal longitudinal data. By definition, an n -dimensional copula $C(u_1, \dots, u_n)$ is a cumulative distribution function with uniform marginals [17–19]. If $F^{(j)}(y_j)$ is the distribution of a continuous univariate random variable Y_j ($j=1, \dots, n$), then $C(F^{(1)}(y_1), \dots, F^{(n)}(y_n))$ is a multivariate distribution for $\mathbf{Y}=(Y_1, \dots, Y_n)^T$ with marginal distributions $F^{(j)}$ ($j=1, \dots, n$). Conversely, it can be proved that for any multivariate distribution $F(y_1, \dots, y_n)$ with continuous marginal distributions $F^{(j)}$ ($j=1, \dots, n$), there exists a copula $C(u_1, \dots, u_n)$ such that

$$F(y_1, \dots, y_n) = C(F^{(1)}(y_1), \dots, F^{(n)}(y_n))$$

Originally, these results were mainly used to simulate multivariate random variables with given marginal distributions. More recently, they have been considered to model dependent data in subjects like survival analysis [20], genetics [21], risk theory [22] and insurance [23].

The plan of the paper is as follows. In Section 2, we present the study that motivated this research. Section 3 is devoted to notation. A copula [17] model for univariate longitudinal (possibly) non-normal data is proposed in Section 4 where the normal case with a structured covariance matrix is described in detail. An extension to non-normal data is then derived. A joint model for all the series which accounts for the dependence between all the responses is described in Section 5. We conclude the paper by the analysis of the motivating study in Section 6 and by a discussion in Section 7.

2. THE MOTIVATING STUDY

The goal of this dose titration study is to learn about the haemodynamic effect of a new antidepressant. The considered compound, an inhibitor of norepinephrine and serotonin uptake, is indicated in the treatment of depression.

A subject blind placebo-controlled randomized study was performed on 12 volunteers (six males and six females) to assess the safety, pharmacokinetics and pharmacodynamics of this drug during a dose escalation phase. One male and one female were on the placebo arm, the other ten volunteers being on the treatment arm. For them, the dose administered was progressively incremented: 20 mg twice daily (b.i.d.) on days 1–2; 40 mg b.i.d. on days 3–8; 60 mg b.i.d. on days 9–14; 80 mg b.i.d. on days 15–19, and 80 mg on the morning of day 20. An abrupt treatment discontinuation was performed on day 20 and monitored up to day 22. The administration of the drug occurred in the morning at 8 a.m. and in the evening at 8 p.m.

Our analysis will focus on safety data and more precisely on the assessment of the effect of the drug on heart rate (HR, in beats/min), diastolic (DBP) and systolic blood pressures (SBP) (in mmHg) for the ten subjects in the treatment arm.

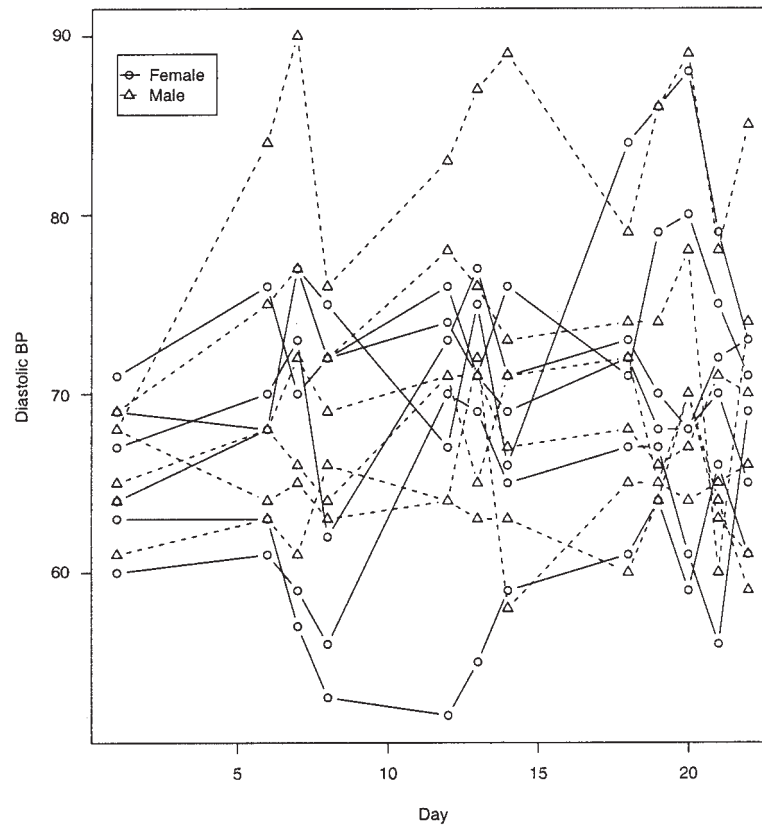


Figure 1. Diastolic BP (in mmHg) 4 hours after the dose administration.

These measurements were taken before the first dose on day 1 and 4 hours after the morning dose on days 6–8, 12–14, 18–22. Thus, 12 repeated measurements were recorded per subject for each of the three response variables. In addition, the drug concentration (in ng/ml) was measured in plasma at the same times. An additional covariate is sex. Plots of the three responses against time can be found in Figures 1, 2 and 3 for the ten subjects.

In the subsequent analysis, we shall use the concentration of the dose found in plasma instead of the administered dose to evaluate the effect of the drug on the three responses. Indeed, plasma concentrations are more reliable measures of exposure than the administered dose during a dose-escalation phase and when steady state conditions are not yet reached. Moreover, a direct association between haemodynamic response and systemic exposure is possible because the two components are measured in the same cardiovascular system.

3. NOTATION

Let $\{t_j: j=1, \dots, J\}$ be the J (possibly unequally spaced) measurement times and i ($i=1, \dots, I=10$) index the I subjects in the study.

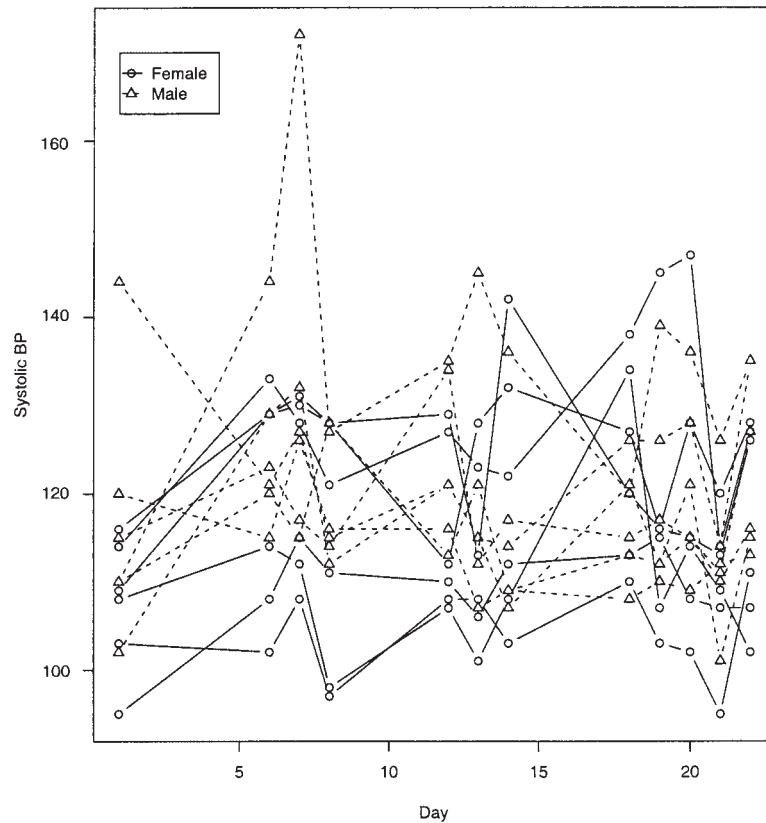


Figure 2. Systolic BP (in mmHg) 4 hours after the dose administration.

For any subject i at time t_j , suppose that we measure R responses $(y_{ij}^{(r)})_{r=1,\dots,R}$. In most of the text, the subject subscript i and the response superscript r will be dropped to simplify the equations. Unless explicitly stated, equations will always refer to any of the R responses observed on any given subject at the indicated time point. This will apply in most of Section 4. The considered response will usually be indicated in Section 5 to stress that parameters are not necessarily shared by the R marginal models for the responses.

4. A STATISTICAL MODEL FOR UNIVARIATE LONGITUDINAL DATA

Let us consider each of the R responses separately. The multivariate aspect of the problem will be addressed in a later section. To fix ideas, consider the J measures of response r on subject i . As already mentioned, i and r will usually be dropped to simplify notation.

4.1. The normal model

The procedure that we propose starts from the well known multivariate normal modelling strategy used in repeated measurements theory (see for example reference [24]). If normality

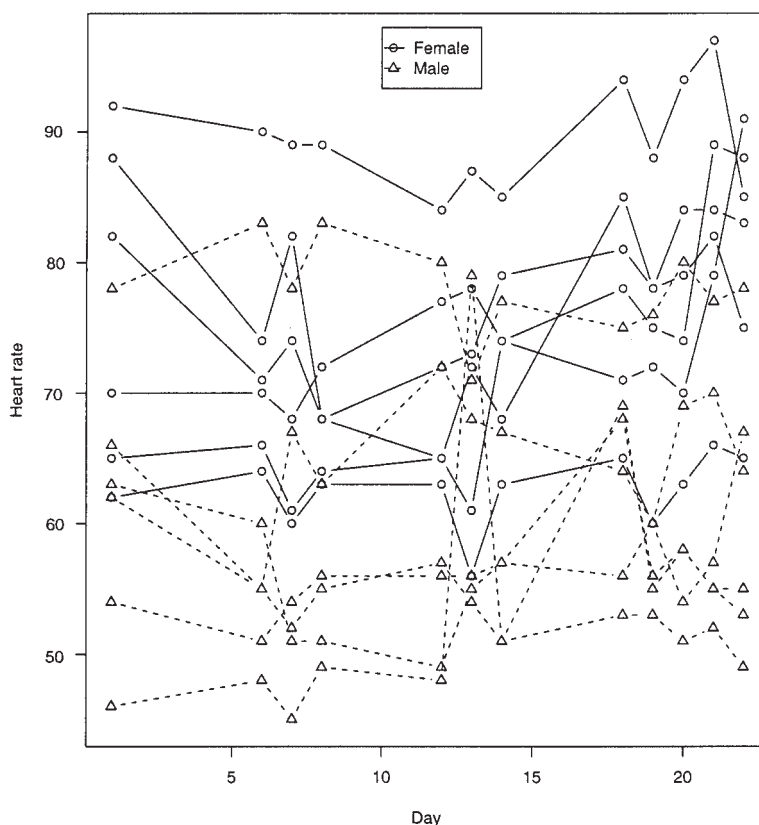


Figure 3. Heart rate (in beats/min) 4 hours after the dose administration.

is believed to be a reasonable hypothesis, a possible model for $\mathbf{Y} = (Y_1, \dots, Y_J)^T$ could be

$$\mathbf{Y} \sim \text{MVN}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu} = (\mu_j)_{j=1, \dots, J}$ is the J -vector of response means and $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \leq j, k \leq J}$ is the variance-covariance matrix. Note that the subscript (here J) used when denoting distributions and densities just indicates the dimension of the corresponding random variable.

The mean can vary with covariates in a way to be defined. The response variance can also be modelled as a function of covariates. Let us denote by $\boldsymbol{\beta}$ the parameters used to describe the covariates effect on the mean and on the variance. Thus, at time t_j , we have

$$\mu_j = \mu(\mathbf{x}_j; \boldsymbol{\beta}); \quad \sigma_{jj} = \sigma_j^2 = \sigma^2(\mathbf{x}_j; \boldsymbol{\beta})$$

where $\mathbf{x}_j$ are the measured covariates at that time for the considered subject.

Two potential sources of dependence in repeated measures are considered:

- (i) *Individual heterogeneity*: measurements of a response realized on the same individual tend to be more alike than across individuals.

- (ii) *Serial dependence*: measurements of a response realized on the same individual at close time points tend to be more closely related than measurements made at more distant times.

The correlation matrix $\mathbf{R}=(\rho_{jk})_{1 \leq j,k \leq J}$ that we assume identical for all subjects is parameterized to account for these two types of dependence in longitudinal data.

A possible choice for $\mathbf{R}=\mathbf{R}(\rho_h, \rho_s)$ is

$$\rho_{jk} = \rho_{jk}(\rho_h, \rho_s) = \rho_h + (1 - \rho_h)\rho_s^{|t_k - t_j|} \quad \text{for } k \neq l \quad (1)$$

where $|\rho_h| \leq 1$ and $|\rho_s| \leq 1$ with the additional constraint that $\mathbf{R}(\rho_h, \rho_s)$ must be positive definite. We have $\rho_{jk} \rightarrow 1$ as $|t_k - t_j| \rightarrow 0$ and $\rho_{jk} \rightarrow \rho_h$ as $|t_k - t_j| \rightarrow \infty$. Thus, as the time interval between two measurements increases, a residual *intra-unit correlation* ρ_h due to individual heterogeneity remains. An extra correlation term $(1 - \rho_h)\rho_s^{\Delta t}$ arises when the time interval Δt is small.

Note that the variance-covariance matrix Σ is a consequence of the chosen models for the response variance and for the correlation matrix as

$$\begin{aligned} \sigma_{jj} &= \sigma_j^2 = \sigma^2(\mathbf{x}_j; \boldsymbol{\beta}) \\ \sigma_{jk} &= \rho_{jk}(\rho_h, \rho_s) \sigma(\mathbf{x}_j; \boldsymbol{\beta}) \sigma(\mathbf{x}_k; \boldsymbol{\beta}) \end{aligned}$$

Estimators of the regression parameters $\boldsymbol{\beta}$ and of the correlation parameters (ρ_h, ρ_s) can be obtained by maximizing the likelihood

$$L(\boldsymbol{\theta}) = \prod_i L_i(\boldsymbol{\theta}) = \prod_i \phi_J(y_{i1} - \mu_{i1}, \dots, y_{iJ} - \mu_{iJ} | \Sigma_i) \quad (2)$$

where:

- (i) $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \rho_h, \rho_s)^T$;
- (ii) $L_i(\boldsymbol{\theta})$ is the likelihood contribution due to subject i ;
- (iii) $\mu_{ij} = \mu(\mathbf{x}_{ij}; \boldsymbol{\beta})$ and $\Sigma_i = \Sigma_i(\boldsymbol{\theta})$;
- (iv) $\phi_J(\cdot | \Sigma)$ denotes the density of a normally distributed J -variate with mean vector 0 and variance-covariance matrix Σ .

A multivariate non-normal distribution where the dependence structure can be specified as easily and as clearly as in the normal case is clearly lacking in the statistical literature. This is precisely the type of model that we propose below.

4.2. Rewriting of the multivariate normal model

Let us rewrite the likelihood (2) in such a way that the normality assumption made for the univariate marginal distribution $F_1(y_{ij} | \boldsymbol{\beta})$ (with mean μ_{ij} and variance σ_{ij}^2) of each response at time t_j clearly comes out. Let us drop index i to simplify notation. If $f_1(y_j | \boldsymbol{\beta})$ denotes the corresponding univariate density and $f_J(y_1, \dots, y_J | \boldsymbol{\theta})$ the joint density of the J repeated measurements on a given subject, then the likelihood contribution due to that subject

is

$$\begin{aligned}
 L(\boldsymbol{\theta}) &= f_J(y_1, \dots, y_J | \boldsymbol{\theta}) \\
 &= \phi_J(y_1 - \mu_1, \dots, y_J - \mu_J | \Sigma(\boldsymbol{\theta})) \\
 &= \frac{1}{\prod_{j=1}^J \sigma_j} \phi_J \left(\frac{y_1 - \mu_1}{\sigma_1}, \dots, \frac{y_J - \mu_J}{\sigma_J} | R(\rho_h, \rho_s) \right) \\
 &= K'(F_1(y_1 | \boldsymbol{\beta}), \dots, F_1(y_J | \boldsymbol{\beta}) | R(\rho_h, \rho_s)) \prod_{j=1}^J f_1(y_j | \boldsymbol{\beta})
 \end{aligned} \tag{3}$$

where

$$K'(u_1, \dots, u_J | R) = \frac{\partial^J}{\partial u_1 \dots \partial u_J} K(u_1, \dots, u_J | R)$$

is the derivative of the *normal copula* (see reference [18], p. 140)

$$K(u_1, \dots, u_J | R) = \Phi_J(\Phi_1^{-1}(u_1), \dots, \Phi_1^{-1}(u_n) | R)$$

yielding

$$K'(u_1, \dots, u_J | R) = \frac{\phi_J(\Phi_1^{-1}(u_1), \dots, \Phi_1^{-1}(u_n) | R)}{\prod_{j=1}^J \phi_1(\Phi_1^{-1}(u_j))}$$

with:

- (i) $\Phi_1(\cdot)$ and $\phi_1(\cdot)$ denoting, respectively, the distribution and the density of a standardized univariate normal variate.
- (ii) $\Phi_J(\cdot | \Sigma)$ and $\phi_J(\cdot | \Sigma)$ denoting the distribution and the density of a normal J -variate with mean 0 and variance-covariance matrix Σ .

Under independence, that is, under the assumption that $\rho_h = \rho_s = 0$, the first factor in equation (3) (corresponding to the derivative of the copula) is 1 and the likelihood reduces to the product of the (marginal) densities at the J time points.

4.3. A non-normal longitudinal model

We propose to relax the normality hypothesis made for the marginal distribution of the response at each time t_j and to consider instead any continuous distribution that might be found appropriate.

If $F_1(y_j | \boldsymbol{\beta})$ denotes that distribution and $f_1(y_j | \boldsymbol{\beta})$ the corresponding density, then equation (3) determines a multivariate continuous distribution (for the J repeated measurements of the considered response for the chosen subject) with marginal continuous distribution $F_1(y_j | \boldsymbol{\beta})$ at time t_j and dependence parameters ρ_h and ρ_s , zero values yielding independence. Note that, unlike in the multivariate normal model, $\rho_{jk}(\rho_h, \rho_s)$ from equation (1) is not the correlation between Y_j and Y_k anymore, although it still measures the strength of the dependence between

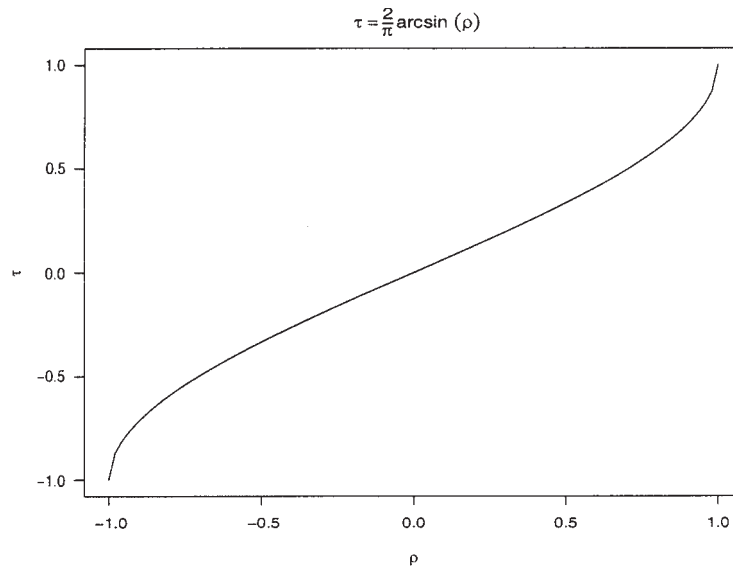


Figure 4. Relationship between the correlation coefficient and Kendall's tau in a bivariate normal model.

these two variables. Indeed, it can be shown (see, for example, reference [18], p. 54) that the Kendall's tau τ_{jk} corresponding to ρ_{jk} is

$$\tau_{jk} = \frac{2}{\pi} \arcsin(\rho_{jk}) \quad (4)$$

As plotted in Figure 4, the dependence between Y_j and Y_k is strictly increasing with ρ_{jk} .

Note that Kendall's tau is not sensitive to monotonic transformations of the data. As we shall consider different marginal distributions for the response in the modelling of our data (see Section 6.1.1), the estimation that we shall obtain for ρ_h and ρ_s should not be too much influenced by the choice made for the marginals.

Other copulas than the normal could be used if judged appropriate. Popular parametric specifications can be found in reference [18], chapter 5.

5. A MODEL FOR MULTIVARIATE LONGITUDINAL DATA

The models in Section 4 treat the R responses separately as if they were independent. In this section, we propose to relate these responses using a copula to connect their conditional (on the past) distributions at each time point.

We start by rewriting the likelihood for each of the R responses as the product of conditional distributions. For the r th response and a given subject, we obtain

$$\begin{aligned} L^{(r)}(\boldsymbol{\theta}^{(r)}) &= f_J^{(r)}(y_1^{(r)}, \dots, y_J^{(r)} | \boldsymbol{\theta}^{(r)}) \\ &= f_1^{(r)}(y_1^{(r)} | \boldsymbol{\theta}^{(r)}) f_1^{(r)}(y_2^{(r)} | \mathcal{F}_2^{(r)}, \boldsymbol{\theta}^{(r)}) \dots f_1^{(r)}(y_J^{(r)} | \mathcal{F}_J^{(r)}, \boldsymbol{\theta}^{(r)}) \end{aligned}$$

with

$$f_1^{(r)}(y_j^{(r)}|\mathcal{F}_{j-}^{(r)}, \boldsymbol{\theta}^{(r)}) = \frac{f_j^{(r)}(y_1^{(r)}, \dots, y_j^{(r)}|\boldsymbol{\theta}^{(r)})}{f_{j-1}^{(r)}(y_1^{(r)}, \dots, y_{j-1}^{(r)}|\boldsymbol{\theta}^{(r)})} \quad (5)$$

where $\mathcal{F}_{j-}^{(r)}$ denotes the history of the r th response (that is, the vector of past responses with all pertinent relationships among them) of the considered subject up to (but not including) time t_j . The expressions for the joint densities in the numerator and denominator of equation (5) are given by equation (3) with marginal distributions $F_j^{(r)}$ and parameters $\boldsymbol{\theta}^{(r)}$ specific to the r th response model.

At each time point t_j for a given subject, we observe $(y_j^{(1)}, \dots, y_j^{(R)})$ for the R responses. The likelihood associated to the R series for this subject is

$$h_R(y_1^{(1)}, \dots, y_1^{(R)}) \prod_{j=2}^J h_R(y_j^{(1)}, \dots, y_j^{(R)}|\mathcal{F}_{j-}) \quad (6)$$

where $h_R(\cdot)$ is the R -variate density associated to the R responses and $\mathcal{F}_{j-}$ denotes the history of the R responses (for the subject) up to (but not including) time t_j .

Let us introduce an assumption on the dependence between the R responses: *we suppose that the r th response at time t_j is, conditionally on its history and conditionally on the other responses at time t_j , independent of the history of these other responses*. In other words, we assume that we can explain the effect of past responses on the r th response by the history of the r th response and by the current values of the other responses.

Under that hypothesis, we have as conditional (on the considered subject's response history) distribution for the R response variables at time t_j

$$\begin{aligned} H_R(Y_j^{(1)} \leq y_j^{(1)}, \dots, Y_j^{(R)} \leq y_j^{(R)}|\mathcal{F}_{j-}) \\ = C_{\boldsymbol{\alpha}}(F_1^{(1)}(y_j^{(1)}|\mathcal{F}_{j-}^{(1)}, \boldsymbol{\theta}^{(1)}), \dots, F_1^{(R)}(y_j^{(R)}|\mathcal{F}_{j-}^{(R)}, \boldsymbol{\theta}^{(R)})) \end{aligned} \quad (7)$$

where $C_{\boldsymbol{\alpha}}(u_1, \dots, u_R)$ is the copula (with dependence parameter vector $\boldsymbol{\alpha}$) relating the conditional (on their respective history) distributions of the R responses at time t_j .

The corresponding conditional density (obtained by taking partial derivatives of the expression in equation (7)) can be used in equation (6) to obtain the contribution to the likelihood for the chosen subject. This likelihood is a function of the vector of parameters $\boldsymbol{\theta}$ where

$$\boldsymbol{\theta}^T = (\boldsymbol{\theta}^{(1)T}, \dots, \boldsymbol{\theta}^{(R)T}, \boldsymbol{\alpha}^T)$$

6. STUDY ANALYSIS

6.1. Univariate responses

Let us first analyse the evolution of diastolic BP, systolic BP and heart rate separately.

6.1.1. Modelling strategy. We have considered different parametric marginal distributions $F_1(y_j|\boldsymbol{\beta})$ for each response. The considered possibilities were the gamma, the Weibull, the inverse Gauss, the normal, the log-normal, the Student and the log-Student.

The chosen location parameter μ_j and the dispersion parameter σ_j of these marginal models were allowed to change with sex and the drug concentration in the plasma. Consider for example the Student distribution. The most general model that we have considered is

$$\begin{aligned}\mu_j &= \lambda_0 + \lambda_1 \text{conc}_j \\ \log(\sigma_j) &= \delta_0 + \Delta\delta_C I_{\{\text{conc}_j \neq 0\}} + \delta_1 \text{conc}_j\end{aligned}\quad (8)$$

with possibly different parameters for males and females. The tail parameter (degree of freedom) v was assumed fixed. Hence, referring to the above notation and taking male as the reference category

$$\boldsymbol{\beta} = (\lambda_{0M}, \Delta\lambda_{0F}, \lambda_{1M}, \Delta\lambda_{1F}, \delta_{0M}, \Delta\delta_{0F}, \Delta\delta_C, \delta_{1M}, \Delta\delta_{1F}, v)^T$$

where, for example

$$\lambda_{0F} = \lambda_{0M} + \Delta\lambda_{0F}$$

is used to obtain the intercept in the location parameter equation for females.

The maximum likelihood estimates (MLE_s) of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \rho_h, \rho_s)^T$ were calculated by a numerical maximization of the likelihood with subject contribution given by equation (3). The function `nlm` in the free software R [25] was used to do this, yielding the required parameter estimates and standard errors. Model selection was realized using the Akaike criterion [26]. Note that there is no allowance for uncertainty in the selection of the marginal distributions for the three responses. We refer to reference [27] for a description of the problem and a possible Bayesian answer.

6.1.2. Results. The selected longitudinal model for each of the three response series is reported in Table I. The number in square brackets gives the change in deviance ($= -2 \log(L)$) when the corresponding parameter:

- (i) is fixed (usually) at zero when it is estimated in the table;
- (ii) is estimated when it is (usually) fixed at zero in the table.

This is given together with standard errors (in brackets) as we cannot discard non-symmetric likelihood profiles for some of the involved parameters; using standard errors in such situations can be misleading. As the AIC was used to do model selection, any parameter causing a decrease in the deviance of at least 2 is maintained in the model.

Note that time does not appear explicitly associated with a regression parameter in the table. Indeed, time is only used to describe serial dependence between the repeated measurements as explained in Section 4.1 and ρ_s was only found necessary to model heart rate profiles. In this dose-escalation study, drug concentration tends to increase with time. For this reason, the effect of time appears indirectly in the model as it is associated with the variation of the drug concentration in plasma.

Table I. MLEs, standard errors (in brackets), deviance change (in square brackets: $-2 \log(L)$ increases when the corresponding parameter is fixed or $-2 \log(L)$ decreases when it is estimated) and likelihood for the selected marginal models for the three series of responses.

	Response								
	Diastolic BP			Systolic BP			Heart rate		
	Distribution								
	Normal			Log-Student			Gamma		
	MLE	SE	Δ Dev	MLE	SE	Δ Dev	MLE	SE	Δ Dev
<i>Location parameters</i>									
λ_{0M}	65.4	(1.11)		4.72	(0.03)		61.1	(3.37)	
$\Delta\lambda_{0F}$	0		[−0.22]	0		[−0.83]	15.0	(4.62)	[+7.27]
λ_{1M}	0.041	(0.013)	[+7.18]	0.00031	(0.00016)	[+3.42]	0		[−0.22]
$\Delta\lambda_{1F}$	0		[−1.85]	0		[−1.45]	0		
<i>Dispersion parameters</i>									
δ_{0M}	1.30	(0.211)		−2.43	(0.14)		−3.65	(0.30)	
$\Delta\delta_{0F}$	0		[−0.01]	0		[−0.03]	0		[−0.40]
$\Delta\delta_C$	0.61	(0.18)	[+7.73]	0		[−0.66]	0		[−0.22]
δ_{1M}	0.0017	(0.0011)	[+2.98]	0		[−0.03]	0		[−0.10]
$\Delta\delta_{1F}$	0		[−0.02]	0			−0.0049	(0.0019)	[+5.80]
<i>Tail parameter</i>									
v	$+\infty$		[−0.67]	7.14	(3.51)	[+4.13]	—		
<i>Dependence parameters</i>									
ρ_h	0.68	(0.06)	[+113.7]	0.56	(0.14)	[+79.58]	0.67	(0.10)	[+32.90]
ρ_s	0		[−0.67]	0		[−0.76]	0.30	(0.10)	[+7.85]
<i>Likelihood</i>									
$-2\log(L)$		846.53			1039.01			937.48	
Parameters		6			5			6	
AIC		858.53			1049.01			949.48	

Let us detail the selected model for each of the three response variables:

- (i) *Diastolic BP*. The selected marginal distribution is Normal. This marginal distribution describes how the diastolic blood pressure of the subjects involved in the study is distributed conditionally on the amount of drug found in their plasma. The considered location and dispersion parameters in equation (8) are the mean and the standard deviation.

No sex effect was detected on the marginal mean. However, the mean DPB increases significantly with the drug concentration.

The dispersion is larger when the drug concentration is not zero, that is, after administration of the drug, and this difference increases with the drug concentration.

There is some evidence of intrasubject association ($\hat{\rho}_h = 0.68 \Rightarrow \hat{\tau}_h = 0.48$) but not of serial dependence. In other words, the DBP measurements tend to be more homogeneous when they are realized on the same individual, but the association remains identical whatever the time interval (considered in the study) separating responses.

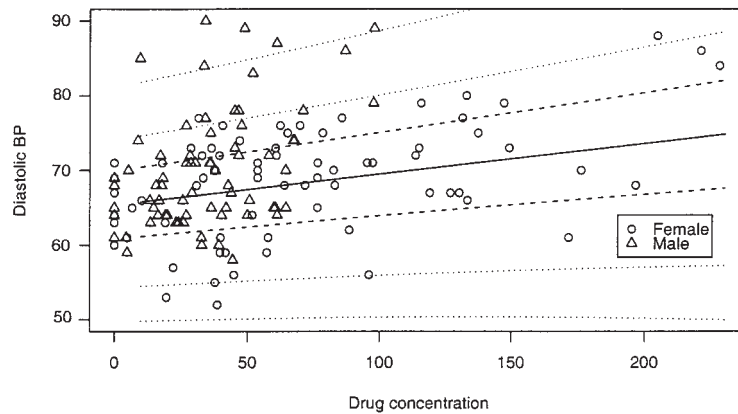


Figure 5. Effect of the drug concentration in the plasma (in ng/ml) on the quantiles 0.01, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.99 of the marginal normal distribution of the (supine) diastolic blood pressure (in mmHg).

The evolution of the fitted (marginal) quantiles for a range of drug concentrations can be found in Figure 5.

- (ii) *Systolic BP*. A log-Student distribution with 7.14 for the degree of freedom is selected. The marginal distribution of the SBP is thus right skewed with a heavier tail than a log-normal, indicating that very large values for the SBP cannot be discarded. Note that a log-normal distribution is selected instead if the very large systolic BP measured for one of the subjects on day 7 (see, for example, Figure 2) is removed. However, we have decided to maintain that observation in the data set as it was really measured. Considering a Student distribution instead of the traditional normal for the log of the systolic BP remains a cautious choice as it allows to make a (parametric) robust inference on the mean response (see reference [28] for a similar argument in another context); outliers mainly affect the degree of freedom.

The considered location and dispersion parameters in equation (8) are the mean of the log SBP and the standard deviation of the log SBP.

There is a small increase in the marginal mean (of the log SBP) with the concentration too, but no significant sex effect.

No significant sex or concentration effect was found on dispersion.

As for the DBP, there is evidence of intrasubject association ($\hat{\rho}_h = 0.56 \Rightarrow \hat{\tau}_h = 0.38$) in the SBP, but not of serial association.

The evolution of the fitted (marginal) quantiles for a range of drug concentrations can be found in Figure 6.

- (iii) *Heart rate*. Heart rate was treated as if it were continuous. The selected distribution is the gamma. The considered parameterization for the gamma density is

$$f(y) = \frac{1}{(\phi\mu)^{\frac{1}{\phi}} \Gamma(\frac{1}{\phi})} x^{\frac{1}{\phi}-1} e^{-\frac{1}{\phi} \frac{x}{\mu}}$$

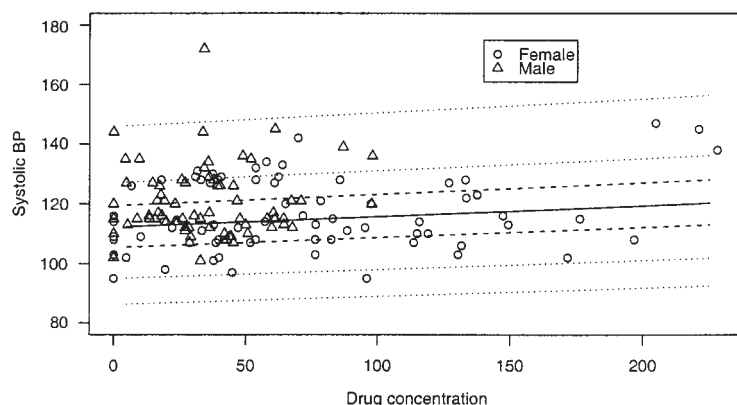


Figure 6. Effect of the drug concentration in the plasma (in ng/ml) on the quantiles 0.01, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.99 of the marginal log-Student distribution of the (supine) systolic blood pressure (in mmHg).

such that

$$E(Y) = \mu \quad \text{and} \quad V(Y) = \phi \mu^2$$

The considered location and dispersion parameters in equation (8) are the mean μ and the parameter ϕ of the gamma conditional distribution for the heart rate.

The marginal mean heart rate is significantly smaller for men than for women. It is not significantly related to the drug concentration.

However, the scale parameter (which is estimated to be $0.026 (=e^{\hat{\delta}_{0M}})$ when no treatment is given) tends to decrease ($\hat{\delta}_{1M} = 0$, $\hat{\delta}_{1F} = -0.0049$) for females (but not for males) as the drug concentration increases.

There is evidence of an intrasubject association ($\hat{\rho}_h = 0.67 \Rightarrow \hat{\tau}_h = 0.46$). Serial dependence was also significant ($\hat{\rho}_s = 0.30$). The Kendall's tau measuring the association between two heart rate measurements on the same subject is thus at least equal to $\hat{\tau}_h$. The necessity to keep the serial association parameter in the model contrasts with the results found for the DBP and the SBP.

The evolution of the fitted (marginal) quantiles for a range of drug concentrations can be found in Figure 7.

6.2. Multivariate responses

The use of copulas allows us to separate the modelling of the three responses and of their dependence. We have already modelled the three series separately in Section 6.1.

We shall use these results with equations (5) and (7) to obtain the conditional distributions required to get the likelihood contribution of subject i as given by equation (6).

The choice of the copula $C_\alpha(u_1, u_2, u_3)$ is arbitrary and should, as much as possible, reflect the dependence structure between the three responses. Here, again, we decided to restrict our

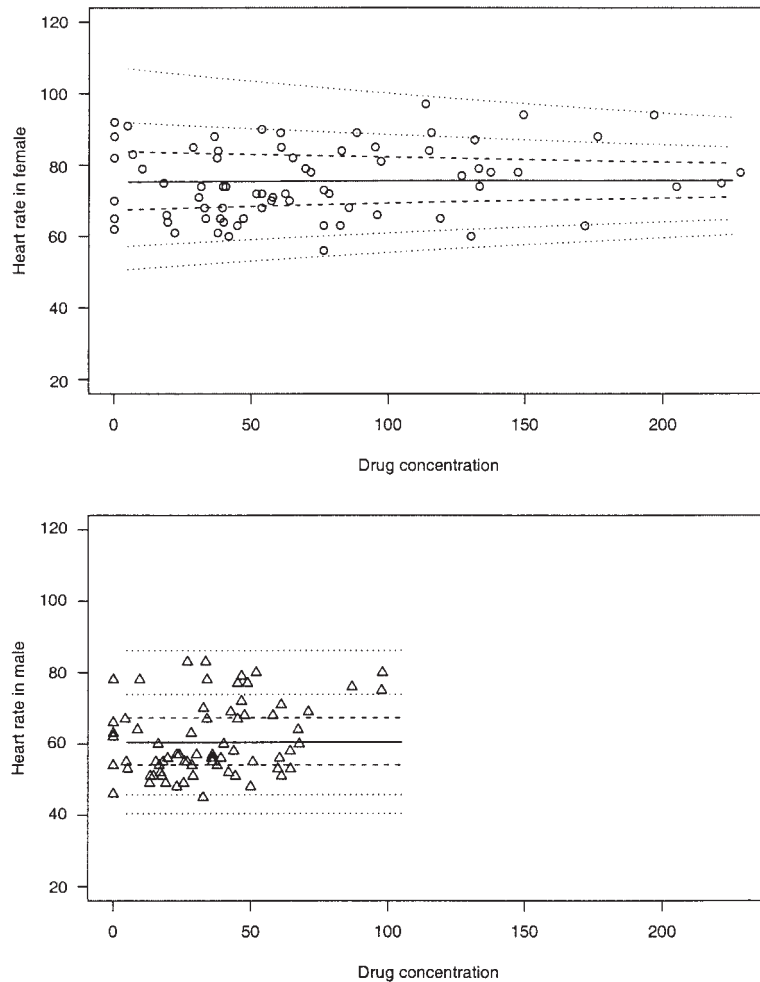


Figure 7. Effect of the drug concentration in the plasma (in ng/ml) on the quantiles 0.01, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.99 of the marginal gamma distribution of the (supine) heart rate (in beats/min).

choice to the normal copula as the dependence structure can easily be specified through the variance-covariance matrix. The dependence between any two of the three responses measured by a parameter ρ (with $|\rho| \leq 1$) is again related to Kendall's tau by equation (4).

Of course, other three-variate copulas could be used without difficulty if judged necessary. We refer the interested reader to references [18] and [19] for a review of useful copulas and of their properties.

Thus, we have

$$C_{\alpha}(u_1, u_2, u_3) = \Phi_3(\Phi_1^{-1}(u_1), \Phi_1^{-1}(u_2), \Phi_1^{-1}(u_3) | \mathbf{R}(\alpha))$$

Table II. MLEs (standard errors) of between-responses dependence parameters and likelihood for the full multivariate model.

	Dependence parameters					
	HR-SBP		HR-DBP		SBP-DBP	
	MLE	SE	MLE	SE	MLE	SE
κ_0^M	0.090	(0.795)	2.346	(0.642)	0.273	(0.959)
$\Delta\kappa_C^M$	0.115	(0.827)	-1.905	(0.682)	1.177	(0.981)
κ_0^F	-0.481	(0.805)	1.415	(0.726)	0.686	(0.768)
$\Delta\kappa_C^F$	0.121	(0.843)	-1.451	(0.768)	-0.003	(0.803)
Likelihood						
$-2 \log(L)$					2764.86	
Marginal parameters					17	
Dependence parameters					12	
AIC					2822.86	

where:

- (i) $\alpha^T = (\alpha^{DS}, \alpha^{DH}, \alpha^{SH})$;
- (ii) $\Phi_3(\cdot | \mathbf{R}(\alpha))$ denotes the three-variate normal distribution with mean 0 and covariance matrix

$$\mathbf{R}(\alpha) = \begin{pmatrix} 1 & \alpha^{DS} & \alpha^{DH} \\ \alpha^{DS} & 1 & \alpha^{SH} \\ \alpha^{DH} & \alpha^{SH} & 1 \end{pmatrix}$$

These (between responses) dependence parameters were allowed to change with sex and presence of drug in the plasma. If α_{ij} denotes such a parameter for unit i at time t_j , the most general (or *full*) model considered for α_{ij} is

$$\log \left(\frac{1 + \alpha_{ij}}{1 - \alpha_{ij}} \right) = \kappa_0 + \Delta\kappa_C I_{\{\text{conc}_{ij} \neq 0\}} \quad (9)$$

with separate parameters for males and females. Note that $\log(\frac{1+\alpha_{ij}}{1-\alpha_{ij}})$ is a strictly increasing function of α_{ij} , with the same sign as α_{ij} . It tends to $+\infty$ ($-\infty$) when α_{ij} tends to $+1$ (-1). A significantly non-zero value for $\Delta\kappa_C$ in equation (9) indicates a significant drug effect. When it is positive (negative), it indicates that the dependence between the two considered responses significantly increases (decreases) when there is drug in the plasma. Independence occurs when both κ_0 and $\Delta\kappa_C$ are zero.

The MLEs (and standard errors) of the between-response dependence parameters associated with the full model (including four parameters for each of the three possible pairs of responses) are reported in Table II. The model was simplified by setting some of these parameters equal

Table III. Between-responses dependence parameters MLEs ($-2 \log(L)$ increases when the corresponding parameter is fixed, usually at 0, or $-2 \log(L)$ decreases when it is estimated) for the selected multivariate model.

	Dependence parameters								
	HR-SBP			HR-DBP			SBP-DBP		
	MLE	SE	ΔDev	MLE	SE	ΔDev	MLE	SE	ΔDev
κ_0^M	0	—	[−0.12]	1.894	(0.406)	[+9.81]	0	—	[−0.13]
$\Delta\kappa_C^M$	0	—	[−0.21]	−1.671	(0.432)	[+7.80]	1.447	(0.206)	[+33.35]
κ_0^F	0	—	[−0.61]	$=\hat{\kappa}_0^M$		[−0.94]	0.717	(0.209)	[+10.10]
$\Delta\kappa_C^F$	0	—	[−1.27]	$=\Delta\hat{\kappa}_C^M$		[−0.59]	0	—	[−0.01]
Likelihood									
$-2 \log(L)$							2769.34		
Marginal parameters							17		
Dependence parameters							4		
AIC							2811.34		

Table IV. Fitted Kendall's tau for pairs of responses in the selected multivariate model.

Sex	Fitted Kendall's tau			
	Drug in plasma	HR-SBP	HR-DBP	SBP-DBP
Male	No	0	0.53	0
	Yes	0	0.07	0.42
Female	No	0	0.53	0.22
	Yes	0	0.07	0.22

to zero or by constraining others to be the same for males and females. Model selection was again performed using the AIC. The MLEs associated with the selected model can be found in Table III. A number in square brackets gives the change in deviance ($= -2 \log(L)$) when the corresponding parameter:

- (i) is fixed (usually) at zero when estimated in the table;
- (ii) is estimated when fixed in the table.

The fitted Kendall's tau corresponding to the selected multivariate model can be found in Table IV. They were obtained using equations (4), (9) and the estimates in Table III.

We can conclude the following:

- (i) *Heart rate–systolic BP association.* There is no significant association between these two variables.
- (ii) *Heart rate–diastolic BP association.* There is a significant positive association between these two variables with a fitted Kendall's tau equal to:

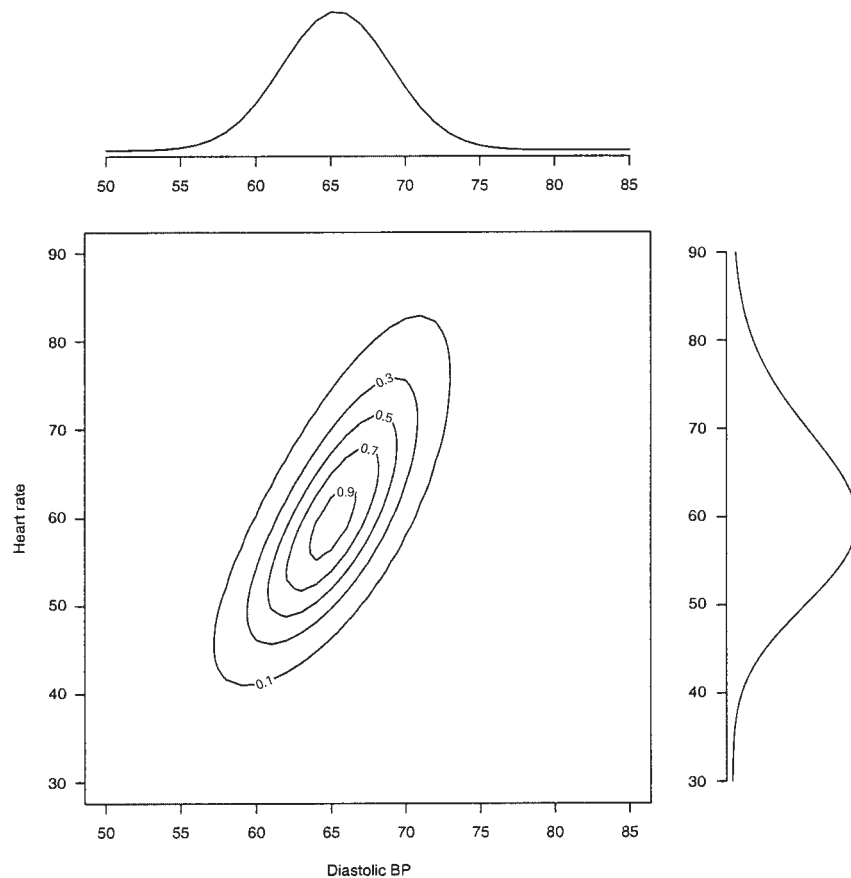


Figure 8. Fitted univariate and bivariate marginal distributions of the heart rate (in beats/min) and of the diastolic BP (in mmHg) for a male without drug in the plasma.

- (a) 0.53 before treatment;
- (b) 0.07 when there is drug in the plasma; although it is small, it is significantly different from zero.

There is no significant effect of sex on the association between HR and DBP. These results are illustrated in Figures 8 and 9. They give the univariate and bivariate marginal distributions of HR and SBP of a male before treatment, in Figure 8, and after treatment (with, say, 35 ng/ml of drug in the plasma), in Figure 9. By comparing the two figures, we clearly see that the DBP is more dispersed after treatment while the heart rate distribution remains nearly unchanged. The positive association between the two responses is visible in both figures, although very moderate on the second. This moderate association (corresponding to a Kendall's tau close to zero) cannot be imputed to the increase in the DBP dispersion as Kendall's tau is insensitive to monotone and hence to scale transformations of the considered variables.

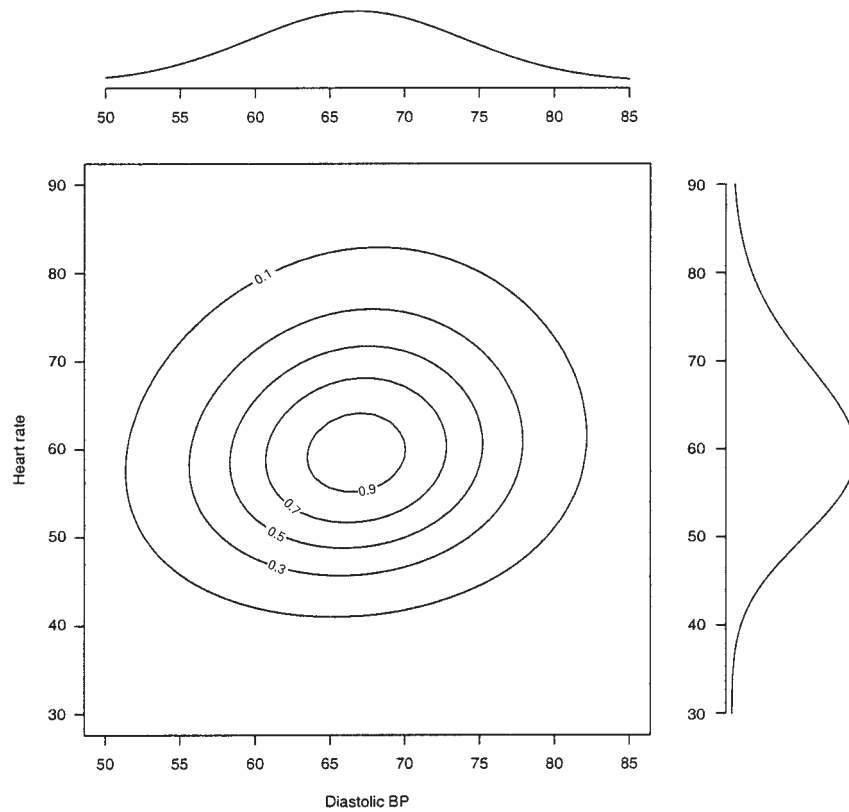


Figure 9. Fitted univariate and bivariate marginal distributions of the heart rate (in beats/min) and of the diastolic BP (in mmHg) for a male with 35 ng/ml of drug in the plasma.

- (iii) *Systolic BP–diastolic BP association.* There is a significant positive association between these two variables with a fitted Kendall's tau equal to:
- (a) 0 and 0.42 for females before and after drug administration, respectively;
 - (b) 0.22 for males; no significant treatment effect on the association parameter was detected.

7. DISCUSSION

In this study, the three responses were treated as if they were multivariate continuous although the heart rate is discrete. This choice was not due to a limitation of the proposed tool. Indeed, copulas can be used whatever the type and the marginal distribution of the involved variables, but when non-continuous variables are involved, one must consider finite differences instead of derivatives of the copula to write down the likelihood. Therefore, the above model can also be used with multivariate longitudinal data of mixed types.

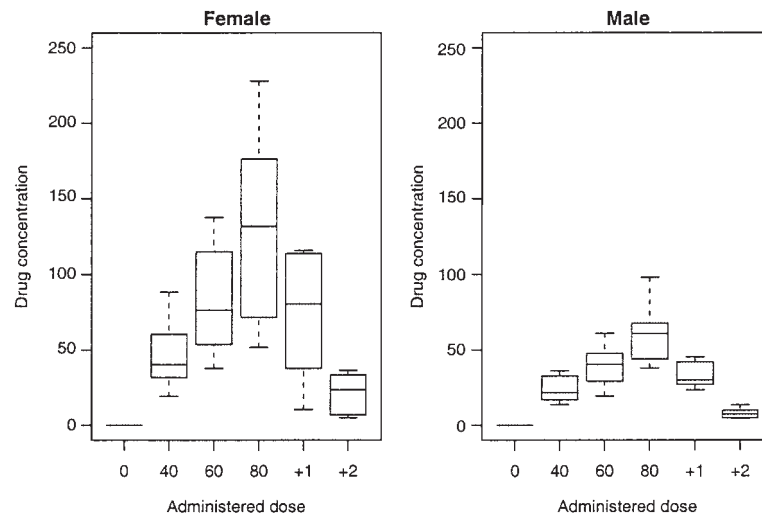


Figure 10. Steady state concentration of the drug in the plasma (in ng/ml) 4 hours after the indicated administered dose (in mg) and then 1 and 2 days after stopping the drug administration.

The normal copula was used throughout the paper for simplicity. Indeed, by specifying two correlation matrices (which should be named more generally *dependence* matrices), we have defined entirely the dependence arising due to the repeated character of our measurements with the first copula and the dependence between the different types of measurements at each time point with the second copula. However, nothing prevents us from using copulas other than the normal one. We refer the interested reader to references [18] and [19] for alternative multivariate copulas. For example, Vandenhende and Lambert (Discussion paper 0025 at <http://www.stat.ucl.ac.be/ISpublication/ISdp.html>, UCL, Belgium) use the multivariate Frank copula to model longitudinal ordinal data.

Except when considering bivariate Archimedean copulas [29], descriptive and graphical tools are clearly lacking to understand the dependence structure in the data and to make a preselection of appropriate copulas. This is the subject of ongoing work that we plan to publish later. For the moment, we select our copula(s) in simple parametric families (indexed by one or two parameters) by taking the one giving the largest log-likelihood.

The estimation of the parameters (from the second copula) describing the dependence between the different variables was made conditionally on the fitted univariate marginal distributions. This uses the argument that continuous margins in a copula model do not depend on the choice of the dependence structure, suggesting that the modelling and the parameter estimations for the margins and the copula can be made separately. One can show (see reference [18], p. 299) that this suggestion can be correct when working with normal copulas.

The drug concentration in the plasma was used as covariate instead of the administered dose for reasons exposed in Section 2. Of course, as soon as the way to administer the medication takes a definitive form and the dosage strength is known, a model relating the responses to the administered dose will become desirable. A pharmacokinetic model relating the drug concentration in the plasma to the administered dose could be developed to become the first building block of the final model, the above model being used as the second block.

Note that sex will certainly play a key role in the pharmacokinetics model as the systemic concentrations were usually larger and more variable in women. This fact is illustrated in Figure 10. Sex did not play a key role as covariate in the models of Sections 6.1 and 6.2 where a drug concentration effect was already present.

REFERENCES

1. Wolfinger R. Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 1986; **1**:205–230.
2. Zeger S, Liang K. Feedback models for discrete and continuous time series. *Statistica Sinica* 1991; **1**:51–64.
3. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:12–22.
4. Jørgensen B, Lundbye-Christensen S, Song PX-K, Sun L. State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics* 1996; **24**:385–402.
5. Olkin I, Tate R. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* 1961; **32**:448–465.
6. Little R, Schluchter M. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 1985; **72**:497–512.
7. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**:1–38.
8. Liu C, Rubin D. Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* 1998; **85**:673–688.
9. Ryan L. Statistical issues in toxicology. *Journal of the American Statistical Association* 2000; **95**:304–308.
10. Fitzmaurice G, Laird N. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* 1995; **90**:845–852.
11. Cox D. The analysis of multivariate binary data. *Applied Statistics* 1972; **21**:113–120.
12. Cox D, Wermuth N. Response models for mixed binary and quantitative variables. *Biometrika* 1992; **79**:441–461.
13. Catalano P, Ryan L. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* 1992; **87**:651–658.
14. Sammel M, Ryan L, Legler J. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B* 1997; **59**:667–678.
15. Regan M, Catalano P. Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics* 1999; **55**:760–768.
16. Ochi Y, Prentice R. Likelihood inference in a correlated probit regression model. *Biometrika* 1984; **71**:531–543.
17. Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'institut de statistique de l'université de Paris* 1959; **8**:229–231.
18. Joe H. *Multivariate Models and Dependence Concepts*. Chapman & Hall: London, 1997.
19. Nelsen R. *An Introduction to Copulas*. Springer-Verlag: New York, 1998.
20. Hougaard P. *Analysis of Multivariate Survival Data*. Springer-Verlag: New York, 2000.
21. Trégouët D, Ducimetière P, Bocquet V, Visvikis S, Soubrier F, Tiet L. A parametric copula model for analysis of familial binary data. *American Journal of Human Genetics* 1999; **64**:886–893.
22. Clemen R, Reilly T. Correlations and copulas for decision and risk analysis. *Management Science* 1999; **45**:208–224.
23. Frees E, Carriere J, Valdez E. Annuity valuation with dependent mortality. *Journal of Risk and Insurance* 1996; **63**:229–261.
24. Potthoff R, Roy S. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; **51**:313–326.
25. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational Graphics and Statistics* 1996; **5**:299–314.
26. Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Inference Theory, Budapest*, Petrov B, Csáki F (eds), 1973; 267–281.
27. Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* 1995; **57**:45–97.
28. Lambert P, Lindsey J. Analysing financial returns using regression models based on non-symmetric stable distributions. *Applied Statistics* 1999; **48**:409–424.
29. Genest C, Rivest L-P. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association* 1993; **88**:1034–1043.