UCL Université catholique de Louvain

INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

ISBA

DISCUSSION PAPER

2016/48

Testing for qualitative heterogeneity: An application to composite endpoints in survival analysis

Oulhaj, A., El Ghouch, A. and R. Holman

Testing for qualitative heterogeneity: An application to composite endpoints in survival analysis

Journal Title
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Abderrahim Oulhaj [1], Anouar El Ghouch [2], Rury R Holman [3]

Abstract

Composite endpoints are frequently used in clinical outcome trials to provide more endpoints, thereby increasing statistical power. A key requirement for a composite endpoint to be meaningful is the absence of the so-called *qualitative heterogeneity* to ensure a valid overall interpretation of any treatment effect identified. Qualitative heterogeneity occurs when individual components of a composite endpoint exhibit differences in the direction of a treatment effect. In this paper, we develop a general statistical method to test for qualitative heterogeneity, that is to test whether a given set of parameters share the same sign. This method is based on the intersection-union principle and, provided that the sample size is large, is valid whatever the model used for parameters estimation. We propose two versions of our testing procedure, one based on a random sampling from a Gaussian distribution and another version based on bootstrapping. Our work covers both the case of completely observed data and the case where some observations are censored which is an important issue in many clinical trials. We evaluated the size and power of our proposed tests by carrying out some extensive Monte Carlo simulations in the case of multivariate time to event data. The simulations were designed under a variety of conditions on dimensionality, censoring rate, sample size and correlation structure. Our testing procedure showed very good performances in terms of statistical power and type I error. The proposed test was applied to a data set from a single-center, randomized, double-blind controlled trial in the area of Alzheimer's disease.

Keywords

Asymptotic test; Bootstrap; Cox model; Interaction-Union principal; Qualitative interaction; Multivariate survival analysis; Right-censoring.

1 Introduction

Multivariate data are usually collected in clinical trials and observational studies to investigate the effect of an intervention or an exposure on a given medical condition. These multivariate data might represent

[1]Institute of public health, College of Medicine & Health Sciences, United Arab Emirates University (UAEU), United Arab Emirates (UAE)
[2]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium
[3]Diabetes Trial Unit (DTU), University of Oxford

Corresponding author:
Abderrahim Oulhaj, Institute of public health, College of Medicine & Health Sciences, United Arab Emirates University (UAEU), United Arab Emirates (UAE).
Email: aoulhaj@uaeu.ac.ae

direct outcomes such as occurrences of different aspects of the medical condition under investigation. This is the case of many clinical trials in cardiovascular disease (CVD) and oncology. In CVD, the effect of an intervention is usually assessed by collecting information on multivariate times to CVD-related events such as time to myocardial infarction (MI), time to stroke and time to cardio-vascular death. In oncology research, the effect of an intervention is commonly assessed by collecting data on times to cancer-related events such as time to cancer progression, time to relapse and time to death from any cause. Multivariate data can also represent some surrogate measures for the medical condition being investigated. This is for instance the case for many clinical trials involving Alzheimer's disease, where multivariate surrogate measures of different aspects of cognitive function such as memory, comprehension, calculation etc are collected and the effect of an intervention is studied. In this framework, a multivariate vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_K)^T$ of possibly correlated components is collected along with a set of explanatory variables Z. We consider the case where Z is a binary variable indicating, for instance, the type of intervention in randomized controlled trials (1 for the active group, 0 for the control or placebo group) or the type of exposure in cohort studies (1 for exposed group, 0 for non-exposed group).

The association between Z and the vector $\boldsymbol{Y}$ is usually investigated by using well established multivariate statistical methods depending on the type of the multivariate distribution of $\boldsymbol{Y}$ given Z. In addition to studying the effects of Z on each component of $\boldsymbol{Y}$, it is also of great interest to investigate if these effects are *qualitatively heterogeneous* in the sense that they do not share the same direction. In the case of a binary covariate Z representing the type of intervention, qualitative heterogeneity holds if there are at least two components of $\boldsymbol{Y}$ such that one shows a beneficial treatment effect but the other shows a harmful treatment effect. One of the many circumstances where qualitative heterogeneity is of interest is the case of composite endpoints. Such endpoints are used commonly in clinical trials to capture the overall effect of an intervention by incorporating several individual components of interest into a single variable. A composite endpoint used frequently in cardiovascular (CV) outcome trials is the 3-point major adverse cardiovascular event (MACE), defined as the time to the first occurrence of CV death, non-fatal myocardial infarction (MI) or non-fatal stroke. In this 3-point MACE example, the composite endpoint is defined as $Y = min(Y_1, Y_2, Y_3)$, where Y_1, Y_2 and Y_3 are the durations of time until the occurrence of the individual components. The absence of qualitative heterogeneity, as defined earlier, is a key requirement for a composite endpoint to be meaningful and to have a valid interpretation. If qualitative heterogeneity is present then interpreting the result of the trial only on the basis of the composite endpoint might be misleading; see, for example, [1]. Testing for the presence of qualitative heterogeneity is therefore of great interest in this case and also in many other research areas.

It is important to note that qualitative heterogeneity deals with the signs of the effects not with their actual values. Another concept related but different from qualitative heterogeneity is the so-called *qualitative or cross-over interaction*; see [2]. Unlike qualitative heterogeneity, defined in a multivariate regression context (i.e., the response $\boldsymbol{Y}$ is a vector of variables), the latter is defined in the context of a univariate multiple regression model and refers to the reversal of the effect of one independent variable on a dependent variable at a certain level of a second independent variable. A common example of qualitative interaction in clinical research is when the active treatment is superior for a subgroup of patients and the alternative treatment or placebo is superior for another subgroup, with the sub-groups defined according to some explanatory variables. Several tests on the detection of qualitative interactions exist in the literature. Gail and Simon [3] published an original paper on how to test for the presence of qualitative interactions in subgroup analysis

using a likelihood ratio test (LRT). An excellent review of qualitative interactions can be found in[2] and various extensions and suggestions for improvement can be found in[4–8]. Although there is an apparent similarity between qualitative heterogeneity and qualitative interaction, the two concepts are completely different because the dependent variable is multivariate in the first case and univariate in the second one. Furthermore, treatment subgroups in qualitative interaction are mutually disjoints as they are defined by the levels of a given covariate whereas in qualitative heterogeneity, no grouping factor is used as the treatment effect is compared according to each component of the vector $\boldsymbol{Y}$ implying all observations to be used in each individual component.

To the best of our knowledge, no formal statistical procedure has been yet developed to test for qualitative heterogeneity. Whilst one might be tempted to use the same statistical tests developed for qualitative interaction, we believe that they cannot be applied in our context as current procedures for testing for qualitative interaction typically assume independent estimates of the effects across the different subgroups. Such independent estimates are generally obtained by separately estimating the treatment effects for each subgroup or center. The aim of this paper is however to develop a general statistical procedure for testing for qualitative homogeneity, taking into account the correlation structure between the different components of the multivariate vector $\boldsymbol{Y}$.

The paper is organized as follows: Section 2 describes the general statistical procedure. In this section, we first provide a clear and formal definition of the concept of qualitative heterogeneity and then we develop a general statistical procedure for testing for its presence. In section 3, we provide an illustration on how to test for qualitative heterogeneity in the case of a composite endpoint and multivariate time to event data. Section 4 demonstrates the properties of our testing procedure through some extensive Monte Carlo simulations and evaluates their sizes and powers for finite and large samples. This was done under different scenarios of dimensionality, censoring and correlations structures between components. In section 5, an application to a real data set from a single-center, randomized, double-blind controlled trial in the area of Alzheimer's disease is provided followed by a concluding discussion in section 6.

2 Qualitative heterogeneity

We consider the association between a multivariate response vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)^T$ of possibly correlated variables and a single binary covariate $Z \in \{0, 1\}$. In medical research, $\boldsymbol{Y}$ might represent several health outcomes of interest measured on the same subject and $Z \in \{0, 1\}$ an indicator of a clinical intervention with 0 and 1 indicating control and treatment arms respectively. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)^T$, be a vector of regression parameters describing the association between Z and each of the components Y_k of the response vector $\boldsymbol{Y}$. Depending on the regression model used, θ_k's can represent any quantity of interest such as population mean differences, median differences, log odds ratios, log hazard ratios or any other parameter quantifying the effect or the association between Z and Y_k. Throughout the paper and without loss of generality, we assume that $\theta_k > 0$ ($\theta_k < 0$) is indicative of a positive (negative) association between Z and Y_k and $\theta_k = 0$ is indicative of no association. According to this, we define qualitative heterogeneity as follows:

Definition 1. *The effects of Z on the vector $\boldsymbol{Y}$ are said to be qualitatively homogeneous if all the components of $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)^T$ have the same sign:*

$$\boldsymbol{\theta} \geq 0 \ \cup \ \boldsymbol{\theta} \leq \mathbf{0} \tag{2.1}$$

Equivalently, the effect are said to be qualitatively heterogeneous if at least two of the components of $\boldsymbol{\theta}$ are of opposite signs or more specifically:

$$\exists\, j \neq k \in \{1, 2, \ldots, K\} \text{ such that } \theta_j \theta_k < 0. \tag{2.2}$$

■

To illustrate the concept, we consider a common example of trials in dementia research where, on each patient, one measures K follow-up responses $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^T$ representing different neuro-psychological tests scores for different areas of the cognitive function such as long and short term memory, calculation, learning, speed and recognition etc, and where patients are randomly assigned to either an active treatment ($Z = 1$) or to control ($Z = 0$). A standard multivariate mean linear regression model characterizing the effect of the treatment Z on the vector of cognitive scores $\boldsymbol{Y}$ can be represented by a set of K linear equations as follows:

$$E(Y_k|Z) = \alpha_k + \theta_k Z, \ k = 1, \ldots, K.$$

The treatment effects on each of the K components of $\boldsymbol{Y}$ are derived as $\theta_k = E(Y_k|Z = 1) - E(Y_k|Z = 0)$, $k = 1, \ldots, K$. In this particular case, qualitative heterogeneity as formulated by (2.2) means that the treatment is beneficial for some areas of the cognitive function but harmful for some others. Note that, in addition to the primary covariate Z, one may adjust for other covariates say $\boldsymbol{X}$ when studying the association between Z and $\boldsymbol{Y}$. In such case, the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)^T$ represent the effect of Z on $\boldsymbol{Y}$ adjusted for $\boldsymbol{X}$ and the definition of qualitative heterogeneity in (2.2) remains the same.

We would like to stress that, in the definition of qualitative heterogeneity in (2.2), the responses Y_k, $k = 1, \ldots, K$ do not necessarily need to be of the same type. They can be a mixture of, for example, continuous, binary and count measurements. This is an interesting and a frequent situation in many clinical studies. In dementia research for example, binary endpoints indicating whether or not a patient performed a certain congestive tasks are typically used to measure attention, visual memory, verbal memory etc. Number of times (count) a participant accurately perform one or more tasks are also important indicators of cognitive skills. In such a case, instead of using a multivariate linear models, one may consider a multivariate generalized linear model of the form:

$$g_k(E(Y_k|Z)) = \alpha_k + \theta_k Z, \ k = 1, \ldots, K,$$

where, for each $k = 1, \ldots, K$, g_k is a given link function (identity, log, logistic, etc) adapted to the type of the component Y_k.
In this work, we are interested to test for the presence of qualitative heterogeneity as defined in (2.2) *i.e.*

$$\begin{aligned} H_0 &: \ \theta_k \leq 0, \ \forall k \in \{1, \ldots, K\} \text{ or } \theta_k \geq 0, \ \forall k \in \{1, \ldots, K\} \\ &\text{against} \\ H_1 &: \ \exists j \neq k \in \{1, \ldots, K\} \text{ such that } \theta_j \theta_k < 0. \end{aligned} \tag{2.3}$$

The null hypothesis H_0 indicates qualitative homogeneity and the alternative hypothesis H_1 indicates qualitative heterogeneity.

Remark 1. *Let consider the case of an univariate multiple regression model, e.g.,*

$$E(Y|Z_1, Z_2 = k) = \alpha_k + \theta_k Z_1,$$

where Y is the univariate response of interest, Z_1 is the main binary covariate ($1 =$ treatment and $0 =$ control) and Z_2 is a grouping factor with K levels (e.g. centers in in multicenter clinical trial). Because, in this context, $\theta_k = E(Y|Z_1 = 1, Z_2 = k) - E(Y|Z_1 = 0, Z_2 = k)$ is the mean difference in Y between treated and placebo in group k, H_1 in (2.3) means in this case qualitative interaction as described in the introduction. This is to say that, depending on the context, (2.3) can describe a hypothesis testing about either qualitative heterogeneity or qualitative interaction. Provided the sample size is large, the testing procedure as described below can be used for both situations even in the case of statistically dependent response(s).

2.1 Testing for qualitative heterogeneity

To start describing the statistical testing procedure, let us define:

$$\eta(\boldsymbol{\theta}) = \max(\theta_1, \theta_2, \ldots, \theta_K, 0).$$

This is a non-negative parameter that vanishes if and only if $\theta_k \leq 0$, $\forall k \in \{1, \ldots, K\}$. Also $\eta(-\boldsymbol{\theta}) = -\min(\theta_1, \theta_2, \ldots, \theta_K, 0)$ is non-negative and vanishes if and only if $\theta_k \geq 0$, $\forall k \in \{1, \ldots, K\}$. So, in terms of η, we can write (2.3) as:

$$H_0 : \ \eta(\boldsymbol{\theta}) = 0 \text{ or } \eta(-\boldsymbol{\theta}) = 0 \text{ against } H_1 : \ \eta(\boldsymbol{\theta}) > 0 \text{ and } \eta(-\boldsymbol{\theta}) > 0. \tag{2.4}$$

or, equivalently as:

$$H_0 : \ H_{0.1} \text{ or } H_{0.2} \ \text{ against } \ H_1 : \ H_{1.1} \text{ and } H_{1.2},$$

where $H_{0.1} : \ \eta(\boldsymbol{\theta}) = 0$, $H_{1.1} : \ \eta(\boldsymbol{\theta}) > 0$, $H_{0.2} : \ \eta(-\boldsymbol{\theta}) = 0$, and $H_{1.2} : \ \eta(-\boldsymbol{\theta}) > 0$.

Let $\hat{\boldsymbol{\theta}}$ be a consistent estimator of the unknown parameter $\boldsymbol{\theta}$ such that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{N} \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.5}$$

where n is the sample size and $\boldsymbol{\Sigma} = [\sigma_{ij}]_{1 \leq i,j \leq K}$ is the (asymptotic) variance-covariance matrix which can also be consistently estimated, say, by $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}]_{1 \leq i,j \leq K}$. To guarantee the asymptotic normality of $\hat{\boldsymbol{\theta}}$, one may specify a parametric form for the joint distribution of $\boldsymbol{Y}$ conditional to Z and apply the classical maximum likelihood theory. However, specifying a joint distribution can be very difficult and complicated especially in the case where the vector $\boldsymbol{Y}$ is a mixture of continuous, discrete and binary outcomes. Various estimating approaches avoiding full specification of the joint distribution can be found in the statistical literature. Among them we cite the marginal pseudo-likelihood approach of[9] which yields to consistent and asymptotically normal estimators, even when the dependence structure between the components of $\boldsymbol{Y}$ is mis-specified. A similar approach for the specific case of multivariate incomplete failure time data was proposed by[10]. This approach will be discussed in more details in section 3. Other widely used methods

include latent variable modeling and general estimating equation. More about this subject can be found in, e.g.[11] and[12], and the references therein.

If the $\hat{\theta}_k$'s were statistically independent and Normaly distributed then one could apply, for example, the likelihood ratio test for qualitative interaction in[3], see also[13] and[2], to test the null hypothesis in (2.4). However, the independence assumption is clearly not realistic in our setting. We propose testing procedures for both $H_{0.1}$ versus $H_{1.1}$ and $H_{0.2}$ versus $H_{1.2}$ and then we use the intersection-union principle to test their union.

So, let us first focus on testing $H_{0.1}$ against $H_{1.1}$. Note that $H_{0.1}$ is a joint hypothesis with multiple inequality constraints given by $\theta_k \leq 0,\ k = 1, \ldots, K$. We consider the approach based on the work of[14] and[15]. Although their framework is completely different from ours (the authors focused on comparing the predictive ability of several regression models in the context of time series analysis) the testing problem is the same. Before discussing this method, let us first introduce a statistic that will play a central role in our testing procedure. Suggested by Hansen[15], this statistic is defined by :

$$T_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \max\left(\sqrt{n}\frac{\hat{\theta}_1 - \max(\theta_1, 0)}{\hat{w}_1}, \ldots, \sqrt{n}\frac{\hat{\theta}_K - \max(\theta_K, 0)}{\hat{w}_K}, 0\right) \tag{2.6}$$

where $\hat{w}_k^2$ can be any consistent estimator of σ_{kk}, the asymptotic variance of $\sqrt{n}\hat{\theta}_k$. For the validity of the test, this studentization is not needed but it leads to a more powerful test. More generally, one can take $\hat{w}_k$ to be any statistics that converge in probability to some positive constant w_k, for $k = 1, \ldots, K$. This might be useful in situations where the individual hypotheses, $\theta_k \leq 0,\ k = 1, \ldots, K$ are not considered equally important. From the proof of Theorem 1 in[15], we have that:

$$T_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \varphi(A_{\boldsymbol{\theta}}^{+}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \boldsymbol{w}_n) + o_p(1), \tag{2.7}$$

where $\varphi(\boldsymbol{x}, \boldsymbol{y}) = \max(x_1/y_1, \ldots, x_K/y_K, 0)$, $\boldsymbol{w}_n = (w_1, \ldots, w_K)^T$, and $A_{\boldsymbol{\theta}}^{+} = diag(I(\theta_k \geq 0)_{1\leq k\leq K})$; i.e. the $K \times K$ diagonal matrix with the $(k,k)th$ element given by $I(\theta_k \geq 0)$. Under $H_{0.1}$, $T_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ in (2.6) reduces to:

$$T_{1n} = \max\left(\frac{\sqrt{n}\hat{\theta}_1}{\hat{w}_1}, \ldots, \frac{\sqrt{n}\hat{\theta}_K}{\hat{w}_K}, 0\right). \tag{2.8}$$

This is the test statistic to be used to test $H_{0.1}$ against $H_{1.1}$. The statistics T_{1n} is similar to the range statistic used by[8] in the context of qualitative interactions. Clearly larger is T_{1n} stronger is the evidence against the null hypothesis $H_{0.1}$. Using (2.5) and (2.7), it easily to see that:

$$T_{1n} \xrightarrow{d} \begin{cases} \varphi(\tilde{\boldsymbol{N}}, \boldsymbol{w}) := \max\left(\frac{\tilde{N}_1}{w_1}, \ldots, \frac{\tilde{N}_K}{w_1}, 0\right), & \text{under } H_{0.1} \\ +\infty, & \text{under } H_{1.1}, \end{cases}$$

where $\tilde{\boldsymbol{N}} := A_{\boldsymbol{\theta}}^0\boldsymbol{N} = (\tilde{N}_1, \ldots, \tilde{N}_K)^T \sim N_K(\boldsymbol{0}, A_{\boldsymbol{\theta}}^0\boldsymbol{\Sigma}A_{\boldsymbol{\theta}}^0)$ with $A_{\boldsymbol{\theta}}^0 = diag(I(\theta_k = 0)_{1\leq k\leq K})$.This implies that the test with the critical region $\mathcal{R}_1 = \left\{\hat{\boldsymbol{\theta}} :\ T_{1n} > c_1\right\}$, where c_1 is such that $\sup_{\boldsymbol{\theta}\leq\boldsymbol{0}} P_{\boldsymbol{\theta}}(\varphi(\tilde{\boldsymbol{N}}, \boldsymbol{w}) > c_1) = \alpha$, is an asymptotically valid $\alpha-$size test for $H_{0.1}$ versus $H_{1.1}$. Note that, from the definition of $A_{\boldsymbol{\theta}}^0$, if $\theta_k < 0$ then the component $\tilde{N}_k$ of $\tilde{\boldsymbol{N}}$ reduces to 0. Because the distribution of $\tilde{\boldsymbol{N}}$ depends on the unknown parameter

vector $\boldsymbol{\theta}$ and becomes degenerate when all the θ_k are strictly negative, getting the exact value of c_1 is impossible. One way to handle this problem is to use the least favorable distribution, that is the asymptotic distribution of T_{1n} under $\boldsymbol{\theta} = \mathbf{0}$, which is the element of the null least favorable to the alternative. In fact, if $\boldsymbol{\theta} = 0$, then $\tilde{\boldsymbol{N}} = \boldsymbol{N} \sim N_K(\mathbf{0}, \boldsymbol{\Sigma})$ and the problem reduces to finding a c_1^0 such that $P(\varphi(\boldsymbol{N}, \boldsymbol{w}) > c_1^0) = \alpha$. Now, one can get an approximate value for c_1^0 by using Monte Carlo (MC) simulations, i.e. by generating a large sample $\boldsymbol{N}^b = (N_1^b, \ldots, N_K^b)^T$, $b = 1, \ldots, B$, from $N_K(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, and solving in c_1^0 the equation:

$$\frac{1}{B}\sum_{b=1}^{B} I(\varphi(\boldsymbol{N}^b, \hat{\boldsymbol{w}}) > c_1^0) = \alpha. \tag{2.9}$$

An approximated Monte Carlo-based p-value for testing $H_{0.1}$ against $H_{1.1}$ using (2.9) is then given by:

$$Pmc_1 = \frac{1}{B}\sum_{b=1}^{B} I(T_{1n}^b > t_{1n}),$$

where $T_{1n}^b = \varphi(\boldsymbol{N}^b, \hat{\boldsymbol{w}})$ and t_{1n} is the observed value of T_{1n}.

There are at least two problems with the MC approach described above: First, using the least favorable configuration may lead to a test with low power; Second, it requires the estimation of the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$ which can be very difficult especially if the components of the vector $\boldsymbol{Y}$ are correlated and of different types.

An alternative and simpler approach is to use the Bootstrap. Let $\hat{\boldsymbol{\theta}}^{*,b}$, $b = 1, \ldots, B$, be a bootstrap sample version of $\hat{\boldsymbol{\theta}}$ constructed using a given bootstrap data generating process. For $b = 1, \ldots, B$, let

$$T_{1n}^{*,b} := T_n(\hat{\boldsymbol{\theta}}^{*,b}, \hat{\boldsymbol{\theta}}) = \max\left(\sqrt{n}\frac{\hat{\theta}_1^{*,b} - \max(\hat{\theta}_1, 0)}{\hat{w}_1}, \ldots, \sqrt{n}\frac{\hat{\theta}_K^{*,b} - \max(\hat{\theta}_K, 0)}{\hat{w}_K}, 0\right)$$

be the recentered bootstrap statistics. The asymptotic bootstrap p-value is then given by:

$$Pbt_1 = \frac{1}{B}\sum_{b=1}^{B} I(T_{1n}^{*,b} > t_{1n}).$$

To test the null hypothesis $H_{0.2}$ against $H_{1.2}$, we follow a similar approach as when testing $H_{0.1}$ against $H_{1.1}$ above. We consider the test statistic

$$T_{2n} = -\min\left(\frac{\sqrt{n}\hat{\theta}_1}{\hat{w}_1}, \ldots, \frac{\sqrt{n}\hat{\theta}_K}{\hat{w}_K}, 0\right),$$

which is nothing but the value, under $H_{0.2}$, of

$$\begin{aligned} T_n(-\hat{\boldsymbol{\theta}}, -\boldsymbol{\theta}) &= -\min\left(\sqrt{n}\frac{\hat{\theta}_1 - \min(\theta_1, 0)}{\hat{w}_1}, \ldots, \sqrt{n}\frac{\hat{\theta}_K - \min(\theta_K, 0)}{\hat{w}_K}, 0\right) \\ &= \varphi(-A_{\boldsymbol{\theta}}^{-}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \boldsymbol{w}_n) + o_p(1), \end{aligned}$$

where, in the last equality, $A_{\boldsymbol{\theta}}^{-} = diag(I(\theta_k \leq 0)_{1\leq k\leq K})$. Using the same argument as above we have:

$$T_{2n} \xrightarrow{d} \begin{cases} \varphi(-A_{\boldsymbol{\theta}}^{0}\tilde{\boldsymbol{N}}, \boldsymbol{w}) =_d \varphi(A_{\boldsymbol{\theta}}^{0}\tilde{\boldsymbol{N}}, \boldsymbol{w}), & \text{under } H_{0.2} \\ +\infty, \quad \text{under } H_{1.2}. \end{cases}$$

The critical region for testing $H_{0.2}$ against $H_{1.2}$, based on least-favorable configuration ($\boldsymbol{\theta} = 0$), is $\mathcal{R}_2 = \left\{\hat{\boldsymbol{\theta}} : T_{2n} > c_1^0\right\}$, where c_1^0 is given by (2.9). Corresponding approximate $p-$values can be obtained either by Monte Carlo or by bootstrap as follow :

$$Pmc_2 = \frac{1}{B}\sum_{b=1}^{B} I(T_{1n}^{b} > t_{2n}) \quad \text{and } Pbt_2 = \frac{1}{B}\sum_{b=1}^{B} I(T_{2n}^{*,b} > t_{2n}),$$

where t_{2n} is the observed value of T_{2n} and $T_{2n}^{*,b} = T_n(-\hat{\boldsymbol{\theta}}^{*,b}, -\hat{\boldsymbol{\theta}})$.

Now we are ready to test the hypothesis of interest (2.3). Our approach is based on the intersection-union principle according to which one reject the global hypothesis $H_0 : H_{0.1} \cup H_{0.2}$ in favor of $H_1 : H_{1.1} \cap H_{1.2}$ if and only if all the individuals hypothesis $H_{0.l}$, $l = 1, 2$, are rejected. According to this principle, the critical region for the above test is given by:

$$\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_1 = \left\{\hat{\boldsymbol{\theta}} : T_n := \min(T_{1n}, T_{2n}) > c_1^0\right\}.$$

and its corresponding $p-$value is obtained by taking the minimum of the $p-$values of the individual hypothesis :

$$Pmc = \min(Pmc_1, Pmc_2), \tag{2.10}$$

for Monte Carlo-based test, and

$$Pbt = \min(Pbt_1, Pbt_2), \tag{2.11}$$

for the bootstrap test. For more details about statistical properties of interaction-union tests, see e.g. [16] and Section 5.3 of Silvapulle and Sen (2005).

3 Illustration: Composite endpoints with time to event data

Testing for qualitative heterogeneity using our procedure could prove useful in many areas. In particular, it can be used to test for qualitative heterogeneity among individual components of a composite endpoint. A composite endpoint is defined as a clinically relevant endpoint constructed from combining other clinically relevant endpoints, termed individual components, into a single variable and is usually used to capture the overall effect of a therapeutic intervention. For more details regarding the definition and motivations for using composite endpoints, especially in time to event analysis; see, e.g., [1]. One of the main clinical assumptions for a composite endpoint to be valid is the absence of qualitative heterogeneity meaning that individual components of the composite endpoint should all share a common direction for the treatment effect; see, e.g., [17].

Current practice in clinical trials utilizing composite endpoints for time-to-event data is generally to power the study according to the primary composite endpoint, and then to investigate the significance of the overall clinical benefit based on the composite endpoint by using statistical tools such as the log

rank test or the Cox proportional hazard model. To investigate for qualitative heterogeneity, separate assessments of treatment effects on each of the individual components are sometimes provided and their corresponding confidence intervals computed and compared. To illustrate this, let us consider the 3-point MACE endpoint often used in CV trials with the treatment effect measured through the hazard ratio using a Cox proportional hazard model. In this case, 95% confidence intervals for hazard ratios are calculated for each individual components of the 3-point MACE composite endpoint *i.e.* CV death, non-fatal myocardial infarction and non-fatal stroke. Qualitative heterogeneity is assumed if there are at least two hazard ratios confidence intervals where one has an upper bound ¡ 1 and the other has a lower bound ¿ 1. The conclusion in such a situation, is that it is difficult to interpret the results based on the composite endpoint alone even if deemed favorable for the intervention under consideration. Otherwise, qualitative homogeneity is assumed and the result for the composite endpoint is considered correctly interpretable. We believe that this way of assessing qualitative heterogeneity is not statistically appropriate because, among other reasons, it inflates the family-wise error rate and it does not take into consideration the correlation structure between individual components of the composite endpoint. This is particularly important as the individual components of the composite endpoint are often highly correlated especially if they represent different aspects of the medical condition under investigation.

In this section, we provide an appropriate and statistically valid approach to test for qualitative heterogeneity in this particular framework. Let $\{\mathcal{E}_1, \mathcal{E}_2, , \ldots, \mathcal{E}_K\}$ be the set of all pre-specified events of interest (individual components) and let $Y_k : k \in \{1, 2, , \ldots, K\}$ be the actual duration of time from a starting point (*e.g.* randomization) until the occurrence of the event $\mathcal{E}_k$ where the durations might be correlated. The composite endpoint, Y, is defined here as the duration of time until the first occurrence of any of the events $\{\mathcal{E}_1, \mathcal{E}_2, , \ldots, \mathcal{E}_K\}$ *i.e.* $Y = min(Y_1, Y_2, \ldots, Y_K)$ (For $K = 3$, the pre-specified events $(\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3)$ may represent for example the individual components of the MACE composite endpoint). Typically, the observed data might be censored due, for example, to insufficient follow-up or withdrawal. Let C_k be a random variable representing the censoring time that prevent the event $\mathcal{E}_k$ from being observed and Z a binary variable indicating the treatment assignment ($Z = 1$ for treatment arm and $Z = 0$ for control arm). For every subject $i, i = 1, 2, ..., n$, under study instead of (Y_k, Z), one only observe $(T_k, \delta_k, Z), k \in \{1, 2, \ldots, K\}$ where $T_k = min(Y_k, C_k)$ and $\delta_k = \mathbb{1}_{\{Y_k \leq C_k\}}$.

To test for qualitative heterogeneity here, one first need to estimate the effects θ_k of Z on each components Y_k of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_K)^T$ taking into account the correlation structure between the Y_k's. Several methods are available in the literature; see, e.g., [18]. Here we use the marginal modeling approach of Wei et al. [10], see also [19]. These models offer a great flexibility as the dependence structure between components of the vector $\boldsymbol{Y}$ do not needs to be specified in advance. Following Wei et al., we assume different proportional hazard functions for each time to event, namely:

$$h_k(t \mid Z) = h_{0k}(t) exp(\theta_k Z), \; k = 1, \ldots, K \tag{3.12}$$

where h_{0k} is an unspecified baseline hazard function for time Y_k and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)^T$ is the vector of the unknown regression parameters each representing the log hazard ratio for time Y_k. In this case, qualitative heterogeneity holds if two log hazards ratio are of opposite signs and can then be investigated by testing the null hypothesis in (2.3).

The parameter vector $\boldsymbol{\theta}$ is estimated by maximizing the partial likelihood $L(\boldsymbol{\theta} \mid z) = \prod_{k=1}^{K} \prod_{i=1}^{n} L_{ki}(\theta_k|z)$ with

$$L_{ki}(\theta_k|z) = \left(\frac{\exp(\theta_k z_i)}{\sum_{j \in R_k(t_i^k)} \exp(\theta_k z_j)} \right)^{\delta_i^k} \tag{3.13}$$

where $R_k(t_i^k)$ is the risk set for event $\mathcal{E}_k$ at time t_i^k, *i.e.*, the set of individuals who had not experienced the event $\mathcal{E}_k$ nor been censored prior to time t_i^k.

Wei *et al.*[10] showed that, under some mild regularity conditions, the estimate $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ as long as the marginal distribution of Y_k, as given by, (3.12) is correctly specified. They also show that $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and provide a consistent estimate for its asymptotic covariance matrix. Therefore, one can use our general method developed in section 2.1 to test for qualitative heterogeneity in this context. This approach will be investigated in details in the following section.

4 Simulations

We performed some extensive Monte Carlo simulations in order to assess the properties of our suggested tests. The simulations were performed for multivariate time to event data under a variety of conditions on dimensionality, censoring rate, sample size and correlation structure between durations. The first objective of these simulations was to evaluate the size and the power of our proposed testing procedure. The second objective was to compare the performances of our tests with some existing methods, namely the Interval Based Graphical Approach (IBGA) of Pan and Wolfe[4] based on simultaneous confidence intervals and the Gail and Simon Likelihood Ratio Test (LRT)[3]; see[2] for a review. Note that these procedures are not designed for testing qualitative heterogeneity but for testing qualitative interaction instead. As explained earlier, the two concepts are completely different but, from mathematical point of view, they share the same objective, namely testing if the components of a given vector of parameters have the same sign. As it will be shown next, our approach is more robust and more powerful.

4.1 Simulation settings

The data were generated according to the following scheme: First, we draw an *i.i.d.* random sample $\boldsymbol{V}_i = (V_{1i}, \ldots, V_{Ki})$, $i = 1, \ldots, n$, of size n from a K-multivariate normal distribution with $\mathbf{0}$ mean and a variance-covariance matrix with diagonal elements equal to 1. Second, we apply the inverse transformation method to $\boldsymbol{V}_i$ in order to obtain a K-dimensional vector of survival times $\boldsymbol{Y}_i = (Y_{1i}, \ldots, Y_{Ki})$ with marginal cumulative distributions $F_k(t|z) = 1 - S_{0k}(t)^{\exp(\beta_k z)}$, where $z \in \{0, 1\}$ and $S_{0k}(t) = \exp(-(t/\lambda_k)^{p_k})I(t \geq 0)$, i.e., the survival distribution of a $Weibull(p_k, \lambda_k)$ random variable with shape p_k and scale λ_k. The durations Y_{ki} are then generated as $Y_{ki} = F_k^{-1}(\Phi(V_{ki})|z)$, $k = 1, \ldots, K$, where Φ is the cumulative distribution of a standard normal variable. Note that $Y_{ki}|Z = z \sim Weibull(p_k, \lambda_k \exp(-\frac{\beta_k z}{p_k}))$. In this simulation we consider three levels of dependency :

(*i*) *Independence*: $Y_1, \ldots, Y_K$ are mutually independent. This has been set up by making the V_k's mutually independent,
(*ii*) *Positive moderate*: Kendall's tau coefficients between pairs of survival times are all fixed to 0.4. This has been set up by making the Pearson correlation coefficient between V_k and $V_{k'}$, for any $k \neq k'$, to be 0.588,
(*iii*) *Positive high*: Kendall's tau coefficients between pairs of survival times are all fixed to 0.8. This has been set up by making the Pearson correlation coefficient between V_k and $V_{k'}$ to be 0.951.

The censoring variables $C_1, \ldots, C_K$ were independently generated from a $Weibull$ distribution with shape

p_k and scale parameter given by $\lambda_k \exp\left(-\frac{\beta_k z}{p_k}\right)\left(\frac{1-\pi_k(z)}{\pi_k(z)}\right)^{1/p_k}$, where $\pi_k(z) = P(C_k > T_k|Z = z)$ is the desired probability of censoring. We consider three levels of censoring :

(i) *uncensored*: $\pi_k(z) = 0$, for all $(k, z) \in \{1, \ldots, K\} \times \{0, 1\}$,
(ii) *moderate*: $\pi_k(z) = 0.3$, for all $(k, z) \in \{1, \ldots, K\} \times \{0, 1\}$,
(iii) *severe*: $\pi_k(z) = 0.7$, for all $(k, z) \in \{1, \ldots, K\} \times \{0, 1\}$

We investigated the case of $K = 3$ and $K = 6$ and used $\lambda_k = 1$ and $p_k = 2$, for $k = 1, \ldots, K$, in all our simulations. We generated 1000 replicate data sets with total samples sizes $n = 200$ and $n = 800$ with equal allocation to the two treatment groups (50% in the placebo group $\{z = 0\}$ and 50% in the treatment group $\{z = 1\}$). For each scenario, we calculate $t_{ki} = \min(y_{ki}, c_{ki})$ and $\delta_{ki} = I(y_{ki} \leq c_{ki})$ and then we estimate the model parameters θ_k's and the corresponding asymptotic variance-covariance matrix using the marginal modeling approach of Wei *et al.*[10] as described in section 3. This was done in R[20] using the function *coxph* from the package *survival*. After that, we calculate the test statistics T_{1n} and T_{2n} and the bootstrap test statistics T^*_{1n} and T^*_{2n}. Three bootstrap data generating processes were considered in this simulation study:

(i) *The nonparametric conditional bootstrap* ***(NB)***:

a. Using simple random sampling with replacement, draw a bootstrap sample of size $n/2$ from the placebo and the treatment group, respectively.
b. From the resulting bootstrap sample, calculate the bootstrap estimate $\hat{\boldsymbol{\theta}}^* = (\theta^*_1, \ldots, \theta^*_K)^T$ using the marginal method.
c. Repeat the above steps for $B = 3000$ times to obtain the bootstrap estimates $\hat{\boldsymbol{\theta}}^{*(b)}$, $b = 1, \ldots, B$.

(ii) *The semiparametric residual bootstrap* ***(RB)***:

a. Compute the residuals $\hat{\epsilon}_{ki} = \hat{S}_{0k}(t_{ki})\exp(\hat{\theta}_k z_i)$, where $\hat{S}_{0k}$ is the Breslow's estimator of the baseline cumulative survival function S_{0k}.
b. Draw a bootstrap sample of size n from $\{(\hat{\epsilon}_{1i}, \delta_{1i}, \ldots, \hat{\epsilon}_{Ki}, \delta_{Ki})\}^n_{i=1}$. Let $\{(\hat{\epsilon}^*_{1i}, \delta^*_{1i}, \ldots, \hat{\epsilon}^*_{Ki}, \delta^*_{Ki})\}^n_{i=1}$, denote the resulting bootstrap censored residual sample.
c. Calculate the so called probability scale bootstrap times $t^*_{ki} = 1 - (\hat{\epsilon}^*_{ki})^{\exp(-\hat{\theta}_k z_i)}$. Then, from the bootstrap sample $(t^*_{1i}, \delta^*_{1i}, \ldots, t^*_{Ki}, \delta^*_{Ki})$, $i = 1, \ldots, n$, calculate the bootstrap estimate $\hat{\boldsymbol{\theta}}^* = (\beta^*_1, \ldots, \theta^*_K)^T$ using the marginal method.
d. Repeat the above steps for $B = 3000$ times to obtain the bootstrap estimates $\hat{\boldsymbol{\theta}}^{*(b)}$, $b = 1, \ldots, B$.

(iii) *The weighted generalized bootstrap* ***(WB)***:

a. Calculate the weighted partial maximum likelihood estimator

$$\hat{\boldsymbol{\theta}}^* = \arg\max \sum_{k=1}^{K} \sum_{i=1}^{n} \ln(L_{ki}(\theta_k|z))\tilde{w}_i,$$

where L_{ki} is given by (3.13) and $\tilde{w}_i$'s are the so called exchangeable bootstrap weights; see e.g.[21]. In this work, we take $\tilde{w}_i = \frac{w_i}{n^{-1}\sum_{i=1}^n w_i}$, where w_i are *i.i.d.* random variables from an Exponential(1).
b. Repeat the above step for $B = 3000$ times to obtain the bootstrap estimates $\hat{\boldsymbol{\theta}}^{*(b)}$, $b = 1, \ldots, B$.

For more details about bootstrap procedures for censored data see[22] and the references therein. The weighted bootstrap has many advantages compared to the classical nonparametric bootstrap. Among them

we mention the fact that it is more "smooth" and much easier to calculate and valid for many parametric and semiparametric models; see[23] and Section 3.5.1.3 of[24] for more details.

For each $K = 3, 6$, we used four values of the parameter $\boldsymbol{\theta}$ to assess the type I error (under H_0) and the power (under H_1). These values are provided in Table 1. This together with the settings described above (dependency, censoring, and n) lead to a total of 288 scenarios. Due to space limitations, we provide in the following section some selected but representative scenarios.

Table 1. The $\boldsymbol{\theta_i}$'s used in our simulation study to generate the data under H_0 and under H_1.

	Type 1 error (H_0)									Power (H_1)								
	$K=3$			$K=6$						$K=3$			$K=6$					
$\boldsymbol{\theta}_1$	0	-0.5	0	0	0	-0.5	-0.5	0	0	-0.12	0.065	0.25	-0.12	-0.046	0.028	0.102	0.176	0.25
$\boldsymbol{\theta}_2$	0	0	0	0	0	0	0	0	0	-0.24	0.005	0.25	-0.24	-0.142	-0.044	0.054	0.152	0.25
$\boldsymbol{\theta}_3$	0	0.5	0	0	0	0.5	0.5	0	0	-0.36	-0.055	0.25	-0.36	-0.238	-0.116	0.006	0.128	0.25
$\boldsymbol{\theta}_4$	1	1	1	1	1	1	1	1	1	-0.48	-0.115	0.25	-0.48	-0.334	-0.188	-0.042	0.104	0.25

Remark 2. *For each $K = 3, 6$, to generate the data under H_1, the values of $\boldsymbol{\theta}_i$, $i = 1, 2, 3, 4$, were calculated according to the formula :*

$$\theta_{ik} = \lambda\eta_i + (k-1)\frac{(1-\lambda^2\eta_i)}{\lambda(K-1)},\ k = 1, 2, \ldots, K, \tag{4.14}$$

with $\eta_i = -0.03i$ and $\lambda = 4$. Here $\theta_{i1} = min_k\theta_{ik} = \lambda\eta_i < 0$, $\theta_{iK} = \max_k \theta_{ik} = 1/\lambda > 0$ and $\eta_i = min_k(\theta_i)max_k(\theta_i)$. In other words, by (4.14), we simply choose K values equally spaced in $[\lambda\eta_i, 1/\lambda]$. The value of λ and η_i were chosen so that the parameters reflect real situations. For instance, applying formula (4.14) for $K = 3$ we obtain $\boldsymbol{\theta_1} = (-0.12, 0.065, 0.25); \boldsymbol{\theta_2} = (-0.24, 0.005, 0.25); \boldsymbol{\theta_3} = (-0.36, -0.05, 0.25)$ and $\boldsymbol{\theta_4} = (-0.48, -0.11, 0.25)$. This respectively corresponds to a vector of hazard ratios $exp(\boldsymbol{\theta_1}) = (0.89, 1.07, 1.28)$; $exp(\boldsymbol{\theta_2}) = (0.79, 1.00, 1.28)$; $exp(\boldsymbol{\theta_3}) = (0.7, 0.95, 1.28)$ and $exp(\boldsymbol{\theta_4}) = (0.62, 0.89, 1.28)$.

4.2 Simulation results

Type I error and power of our Monte Carlo-based test (MC) and our three bootstrap tests (NB, RB, WB) were estimated empirically from 1000 replications. We also calculate the type I error and the power of the likelihood ratio test (LRT) of Gail and Simon, see[3], and the type I error and the power of the test based on simultaneous confidence intervals Interval (IBGA) of Pan and Wolfe, see[4]. This was done using the function *qualval* from the R package *QualInt*.

For Type I error, the results are summarized in Table 2 and Figures 1-2 given below. Table 2 shows the effect of censoring, dependency, dimension and sample size on the type I error when $\boldsymbol{\theta} = \boldsymbol{\theta}_3$. In Figure 1 we investigate the effect of the level α and Figure 2 shows how the type I error, as function of $\boldsymbol{\theta}$, varies with the sample size n and the dimension K. Depending on the true value of $\boldsymbol{\theta}$, percentage of censoring and correlation, all the studied tests can be either conservative (reject too often) or liberal (accepts too often). The highest rejection probability is attained at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$, see table 1, and the smallest is attained at $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_4$. However, in the majority of the scenarios, we can see that these tests are (asymptotically) valid since the observed Type I error probabilities are always less (or very slightly larger) then the corresponding nominal levels. This is specially true for $n = 800$. For $n = 200$ all the tests tend to be conservative and, globally, IBGA method is the most conservative among the sixths methods. Typically, increasing the censoring percentage or/and increasing the correlation makes the tests more conservative. On the other hand, as the sample size increases, the results become better (Type I error closer to the nominal level). In the independent case, Bootstrap based tests (NB, RB, WB) can result in slightly liberal decisions. This is the opposite of the LR test which become

liberal with highly correlated data. The dimension K has a little or no effect on the results except for LR and IBGA method. In fact, we noticed a decrees on the rejection probability when the number of dependent variables increases from $K = 3$ to $K = 6$. For our testing procedures (MC, MB, Rb and WB), we get a very similar pictures with $K = 3$ and $K = 6$ dependent variables.

For the Power, the results are summarized in Table 3 and Figures 3, 4 and 5 given below. Clearly, the most important factors that affect the power of all the studied tests are sample size and percentage of censoring. As expected, the latter has a negative effect whereas the former has a positive effect. Of course, the power also decrees with the level of the test and can be very small when the sample size is not large enough and/or when censoring is large. Another factor that affect power is the correlation among dependent variables. As this correlation increases, the power decreases. This is especially the case when the small size is note large ($n = 200$). When we compare the different methods, we can see that our WB method (based on the wighted bootstrap), leads almost systematically to the highest power. The other boostarp methods (RB and NB) compare very well with WB. Typically, the MC method leads to the smallest power followed by IBGA and LR. The power function of thees tests seems to be very sensitive to the correlation. Compared to the WB, the power of the LR test is very small when the correlation is high. Regarding the dimension K, we can see again that the results for 3 and 6 responses are very similar. However, in the case of non independence, the gain in the power when using our WB test becomes clearer when we increase the dimension from $K = 3$ to $K = 6$. This can be seen by comparing Figures 3 and 4 with Figure 5.

5 Applications to real data from Alzheimer's disease: The VITACOG study

In this section, the proposed test was applied to a data set from Alzheimer's disease. The data were collected from a single-center, randomized, double-blind controlled trial called VITACOG. The primary objective of this trial was to determine if lowering of the level of a blood marker, called homocysteine (tHcy), with B vitamin supplements (folic acid, vitamins B_6 and B_{12}) over 2 years would slow the rate of brain atrophy in older participants with Mild cognitive impairment (MCI) [25]. MCI is a syndrome defined as "cognitive decline greater than that expected for an individual's age and education level but that does not interfere notably with activities of daily life" [26]. It is considered as a potential precursor to the development of Alzheimer's disease. For more details regarding this study please see [25]. In this study, a total of 168 MCI patients were randomly assigned to either a B-vitamin treatment ($n_1 = 85$) or to a placebo ($n_0 = 83$). The sample size above includes only patients who completed the cranial MRI scans at the start and the end of the study. The primary outcome was the rate of brain atrophy during the first 2 years of treatment denoted here by Y_1. Secondary outcomes were a variety of neuro-psychological test scores reflecting different aspects of the cognitive function. We are interested here to three cognitive scores that are representative of particular cognitive domains: The Hopkins Verbal Learning Test (HVLT-DR) assessing verbal learning and memory (recognition and recall). This score ranges from 0 to 12 and is denoted here by Y_2. The Telephone Interview for Cognitive Status (TICS) ranging from 0 to 40 and denoted by Y_3. Cognitive domains measured by the TICS include orientation, concentration, short-term memory, language, praxis, and mathematical skills. The clinical dementia rating scale (CDR) assessing six domains of cognitive and functional performance such as memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. This score is an ordinal variable with four categories but for illustration and following [27] we used a binary version denoted by Y_4 . So here, for each patient, one observe a vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$ of outcomes where Y_1, Y_2 and Y_3 are treated as continuous variables and Y_4 as a binary one. We also have a binary variable Z representing the treatment assignment ($Z = 1$ if treated with B-vitamins and $Z = 0$

if placebo). The goal is now to test whether the effect of B-vitamin treatment is homogeneous across all components of $\boldsymbol{Y}$ or not. In addition to the weighted bootstrap, we will test for qualitative heterogeneity using the two existing methods mentioned earlier namely the Gail & Simon LRT and the IBGA tests.

Figure 6 shows some aspects of the distribution of $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$ according to treatment status (Z). To estimate the effect of Z on the vector $\boldsymbol{Y}$, we used a linear regression model for the variables Y_1, Y_2 and Y_3 and a logistic regression model for the binary variable Y_4 without explicit specification of the correlation between the four components . The estimate of the effects and their standard errors were $(\hat{\theta}_1 = -0.35, s_1 = 0.11), (\hat{\theta}_2 = -0.69, s_2 = 0.61), (\hat{\theta}_3 = -0.74, s_3 = 0.83)$ and $(\hat{\theta}_4 = -0.46, s_4 = 0.34)$ for respectively Y_1, Y_2, Y_3 and Y_4. From these parameter estimates and looking at figure 6, one can see that B-vitamins treatment effects might be qualitatively homogeneous. To check this, we carried out our qualitative heterogeneity testing procedure as described in section 2. The p-values for testing qualitative heterogeneity were 0.26, 0.34 and 0.2 for respectively the Gail and Simon LRT, IBGA and the WB tests showing no evidence of heterogeneity of treatment effects across the four components.

6 Discussion

Motivated by the composite endpoint framework often used in the context of multivariate survival data, we have developed a statistical testing procedure to check whether the components of a given unknown vector of parameters share the same sign. We defined the concept of qualitative heterogeneity and we demonstrated that this procedure can be used to test for it. We showed the asymptotic validity of our approach and investigated it's finite sample performances via extensive simulations. The simulations were performed in the context of multivariate survival data and under a variety of conditions on dimensionality, censoring rate, sample size and correlation structure. Among the testing procedures that we proposed, overall the weighted bootstrap method showed the best results in terms of power and type I error especially when the dimension increases and/or when the correlation between the components of the multivariate vector increases. Our approach can also be used to test for qualitative interaction instead of current methods which require independent estimates of effects across groups. Finally, we would like to mention that in the case of composite endpoints derived from multivariate survival data, it is often the case that one of the components is an absorbent event. The later means that the occurrence of the event will prevent other individual events from occurring. This is the case for instance of CV death event used in the MACE composite endpoint described earlier. Dealing with qualitative heterogeneity within such a context requires further research.

References

1. Montori VM, Permanyer-Miralda G, Ferreira-González I et al. Validity of composite end points in clinical trials. *Bmj* 2005; 330(7491): 594–596.
2. Yan X and Su X. *Testing for qualitative interaction*, chapter 161. CRC Press, 2005. pp. 1343–1348.
3. Gail M and Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; 41: 361–372.
4. Pan G and Wolfe DA. Test for qualitative interaction of clinical significance. *Statistics in medicine* 1997; 16: 1645–1652.
5. Ciminera JL, Heyse JF, Nguyen HH et al. Tests for qualitative treatment-by-centre interaction using a "pushback" procedure. *Statistics in Medicine* 1993; 12(11): 1033–1045.
6. Kitsche A and Hothorn LA. Testing for qualitative interaction using ratios of treatment differences. *Statistics in medicine* 2014; 33: 1477–1489.
7. Russek-Cohen E and Simon RM. Qualitative interactions in multifactor studies. *Biometrics* 1993; 49(2): 467–477.
8. Piantadosi S and Gail MH. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* 1993; 12(13): 1239–1248.
9. Zhao LP, Prentice RL and Self SG. Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society Series B (Methodological)* 1992; 54: 805–811.
10. Wei LJ, Lin DY and Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association* 1989; 84: 1065–1073.
11. Aerts M, Molenberghs G, Ryan LM et al. *Topics in modelling of clustered data*. CRC Press, 2002.
12. De Leon AR and Chough KC. *Analysis of mixed data: methods & applications*. CRC Press, 2013.
13. Silvapulle MJ. Tests against qualitative interaction: Exact critical values and robust tests. *Biometrics* 2001; 57: 1157–1165.
14. White H. A reality check for data snooping. *Econometrica* 2000; 68: 1097–1126.
15. Hansen PR. A test for superior predictive ability. *Journal of Business & Economic Statistics* 2005; 23: 365–380.
16. Berger RL. *Advances in Statistical Decision Theory and Applications*, chapter Likelihood Ratio Tests and Intersection-Union Tests. Birkhäuser Boston, 1997. pp. 225–237.
17. Freemantle N, Calvert M, Wood J et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *Jama* 2003; 289(19): 2554–2559.
18. Hougaard P. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.
19. Lin D. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine* 1994; 13: 2233–2247.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.
21. Wellner JA and Zhan Y. Bootstrapping z-estimators. Technical report, 1996.
22. Loughin TM. *Bootstrap applications in proportional hazards models*. PhD Thesis, 1993.
23. Cheng G. Moment consistency of the exchangeably weighted bootstrap for semiparametric m-estimation. *Scandinavian Journal of Statistics* 2015; 42: 665–684.
24. Li J and Ma S. *Survival analysis in medicine and genetics*. CRC Press, 2013.
25. Smith AD, Smith SM, De Jager CA et al. Homocysteine-lowering by b vitamins slows the rate of accelerated brain atrophy in mild cognitive impairment: a randomized controlled trial. *PloS one* 2010; 5: e12244.
26. Gauthier S, Reisberg B, Zaudig M et al. Mild cognitive impairment. *The Lancet* 2006; 367(9518): 1262–1270.

27. Oulhaj A, Jernerén F, Refsum H et al. Omega-3 fatty acid status enhances the prevention of cognitive decline by b vitamins in mild cognitive impairment. *Journal of Alzheimer's Disease* 2016; 50(2): 547–557.

Table 2. $100\times$ probability of Type 1 error obtained with $\boldsymbol{\theta} = \boldsymbol{\theta}_3$ (see Table 1) for responses that are independent (I), moderately correlated (M) or highly correlated (H). $\alpha = 5\%$

$cens = 0\%$												
	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	3.0	3.1	1.6	2.9	2.3	1.2	3.3	3.2	4.9	3.5	3.4	2.8
NB	5.6	4.2	2.2	6.1	2.9	1.7	6.3	4.3	5.7	7.2	5.4	3.0
RB	5.0	4.1	1.8	5.8	2.6	1.1	6.4	4.3	5.2	7.1	5.0	3.1
WB	7.1	4.6	2.5	7.1	3.7	2.7	6.6	4.7	6.1	7.7	5.6	3.2
IBGA	5.0	4.0	0.5	3.2	2.0	0.0	5.3	3.9	3.9	4.3	2.9	1.3
LR	4.9	5.0	1.6	3.1	5.4	3.8	5.2	5.9	7.9	3.4	5.7	5.9
$cens = 30\%$												
	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	3.4	1.3	0.6	2.5	1.2	0.4	3.9	4.4	3.0	2.6	4.0	4.0
NB	5.0	3.2	0.9	5.1	2.0	0.5	6.3	5.7	4.0	6.3	5.3	4.3
RB	4.7	2.7	0.9	4.8	1.6	0.3	6.2	5.4	4.1	6.3	5.1	4.5
WB	6.0	3.7	1.3	6.3	2.3	0.8	6.2	6.1	4.5	6.5	5.5	4.6
IBGA	4.5	2.6	0.6	3.0	1.1	0.1	5.0	5.5	3.4	3.9	3.9	2.9
LR	4.7	3.2	1.2	3.1	2.2	1.5	5.5	7.3	5.4	3.0	6.2	6.4
$cens = 70\%$												
	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	2.1	0.5	0.0	1.7	0.5	0.2	2.9	3.7	2.7	4.1	2.7	3.1
NB	2.6	0.6	0.0	2.4	0.9	0.2	5.6	6.6	3.7	7.7	4.3	4.2
RB	2.7	0.7	0.0	2.1	0.8	0.4	5.6	6.4	3.3	8.3	4.0	4.3
WB	3.6	1.0	0.2	3.8	1.7	0.7	6.1	7.0	3.9	8.5	4.7	4.8
IBGA	3.4	1.2	0.0	2.4	0.6	0.2	4.6	5.7	3.2	5.0	2.7	2.1
LR	3.2	1.0	0.1	1.9	0.4	0.3	4.3	6.3	5.7	3.5	4.7	5.7

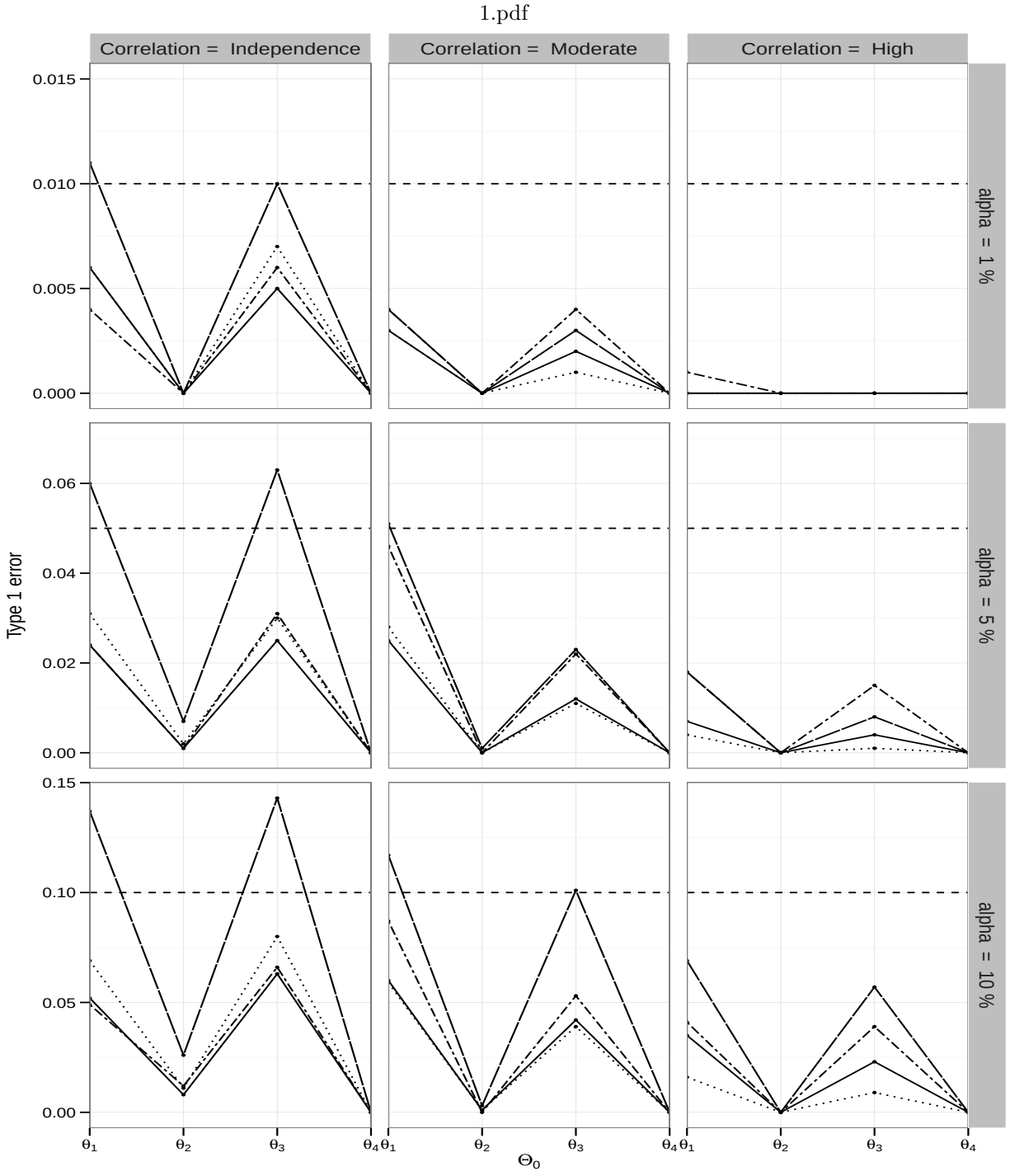

Figure 1. Type I error in the case of six multivariate survival times (K = 6), $n = 200$ and censoring $= 30\%$. Solid line (——) is the MC test, solid dotted line (– – –) is the weighted bootstrap (WB) test, doted line (·····) is the test based on confidence intervals (IBGA) and two-dashed line (-·-·-) is the likelihood ratio test (LR); see Table 1 for the values of $\boldsymbol{\theta}$'s.

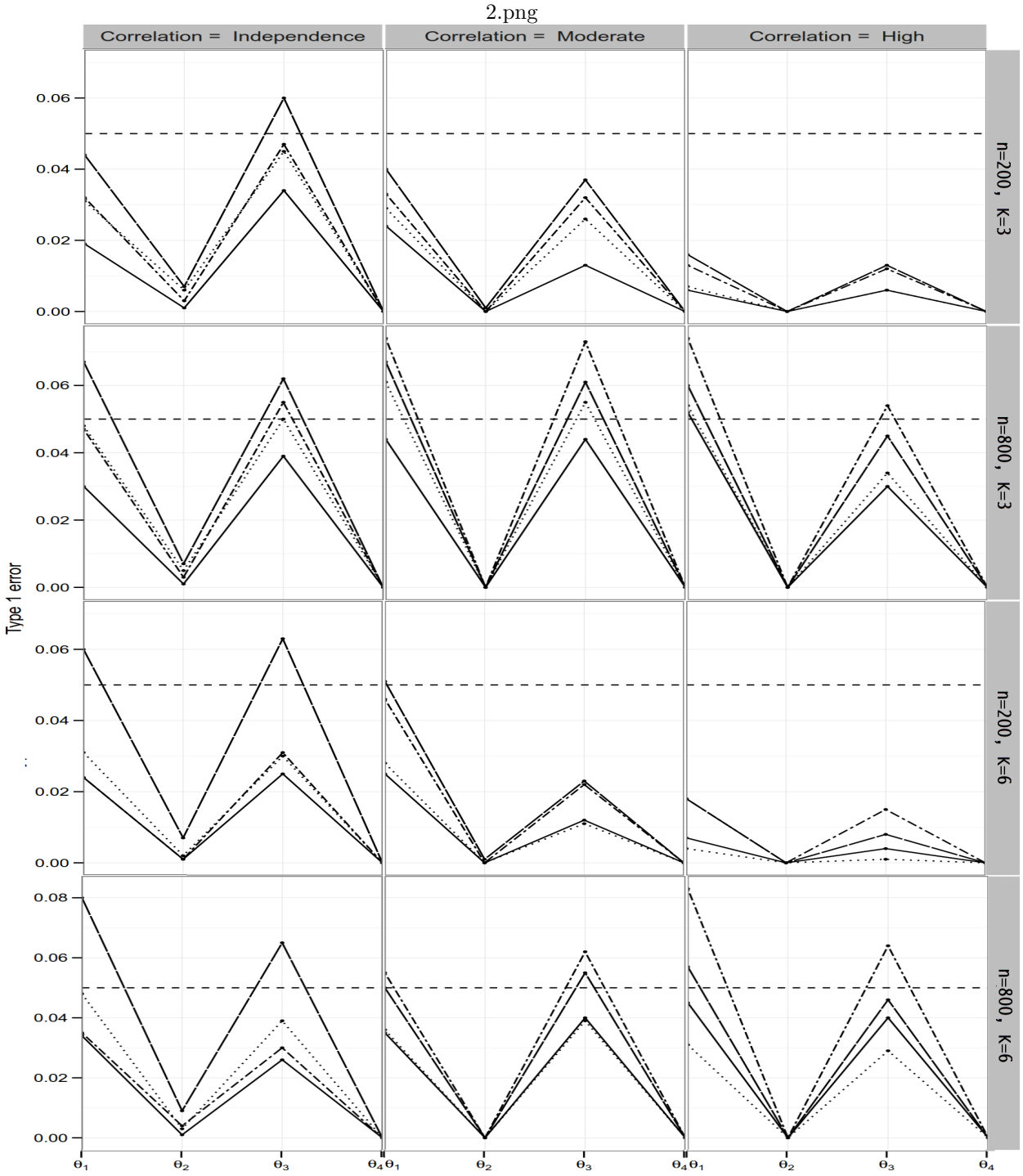

Figure 2. Type I error. Censoring $= 30\%$ and $\alpha = 5\%$. Solid line (——) is the MC test, solid dotted line (– – –) is the weighted bootstrap (WB) test, doted line (·····) is the test based on confidence intervals (IBGA) and two-dashed line (-·-·-) is the likelihood ratio test (LR); see Table 1 for the values of $\boldsymbol{\theta}$'s.

Table 3. $100\times$ Power obtained with $\boldsymbol{\theta} = \boldsymbol{\theta}_3$ (see Table 1) for responses that are independent (I), moderately correlated (M) or highly correlated (H).

$cens = 30\%$, $\alpha = 1\%$

	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	2.6	1.7	0.2	2.1	0.3	0.4	56.7	55.3	55.4	48.6	45.8	48.4
NB	3.7	1.7	0.3	2.0	0.5	0.4	66.2	65.1	64.8	60.1	56.9	59.4
RB	3.6	1.5	0.4	1.8	0.3	0.3	66.0	65.9	66.1	59.6	56.2	57.8
WB	**5.5**	**2.4**	**0.7**	**3.5**	**1.0**	**0.5**	**67.6**	**66.7**	**67.7**	**61.5**	**59.8**	**60.9**
IBGA	4.1	2.3	0.3	2.4	0.3	0.1	62.4	60.5	58.8	53.3	47.1	46.3
LR	4.4	1.3	0.2	2.8	0.3	0.1	59.6	56.6	52.9	56.2	48.8	46.7

$cens = 30\%$, $\alpha = 5\%$

	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	14.5	8.6	5.5	12.9	5.4	3.2	79.4	78.5	81.6	76.3	75.4	79.3
NB	20.8	15.1	9.0	17.8	8.5	4.5	87.5	**87.5**	88.2	84.7	84.9	**87.1**
RB	21.2	16.1	8.0	16.8	8.1	4.4	87.4	86.6	88.2	84.3	84.5	86.5
WB	**23.3**	**17.6**	**10.9**	**20.5**	**11.0**	**7.0**	**87.8**	**87.5**	**89.1**	**85.1**	**85.4**	86.6
IBGA	19.7	13.3	6.2	15.6	5.1	1.4	84.4	82.0	82.7	79.1	75.5	74.4
LR	18.7	10.7	4.9	15.8	4.1	1.6	82.6	79.5	80.1	79.3	73.7	68.9

$cens = 70\%$, $\alpha = 5\%$

	$n = 200$						$n = 800$					
	$K = 3$			$K = 6$			$K = 3$			$K = 6$		
Method \ Corr	I	M	H	I	M	H	I	M	H	I	M	H
MC	3.8	1.4	1.0	2.9	0.6	0.0	31.6	25.5	27.8	30.9	20.1	19.4
NB	5.6	2.1	1.4	3.4	1.2	0.2	43.7	37.1	40.2	43.2	32.3	30.2
RB	4.9	2.1	1.2	3.1	1.1	0.0	43.6	36.8	39.9	42.8	31.5	29.9
WB	**7.1**	**3.0**	**2.2**	**5.7**	**2.1**	**0.5**	**45.7**	**38.5**	**41.1**	**45.6**	**34.0**	**32.2**
IBGA	6.0	2.0	1.4	3.6	0.8	0.0	40.0	31.3	32.7	35.1	20.7	17.2
LR	6.2	2.2	1.5	3.3	0.9	0.0	39.5	28.5	30.3	36.7	24.4	15.9

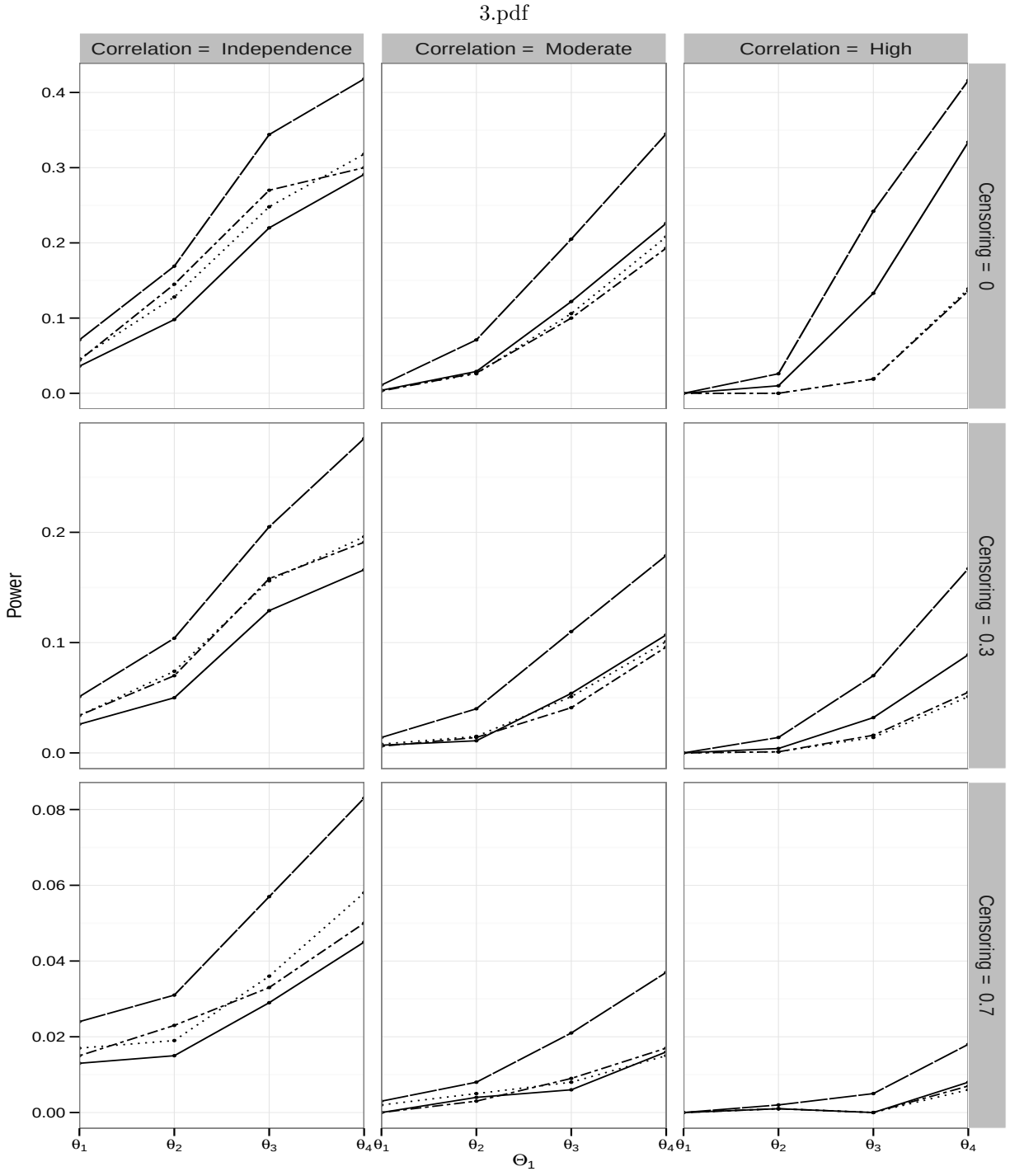

Figure 3. Power function in the case of six multivariate survival times (K = 6), $n = 200$ and $\alpha = 5\%$. Solid line (——) is the MC test, solid dotted line (– – –) is the weighted bootstrap (WB) test, doted line (·····) is the test based on confidence intervals (IBGA) and two-dashed line (-·-·-) is the likelihood ratio test (LR); see Table 1 for the values of $\boldsymbol{\theta}$'s.

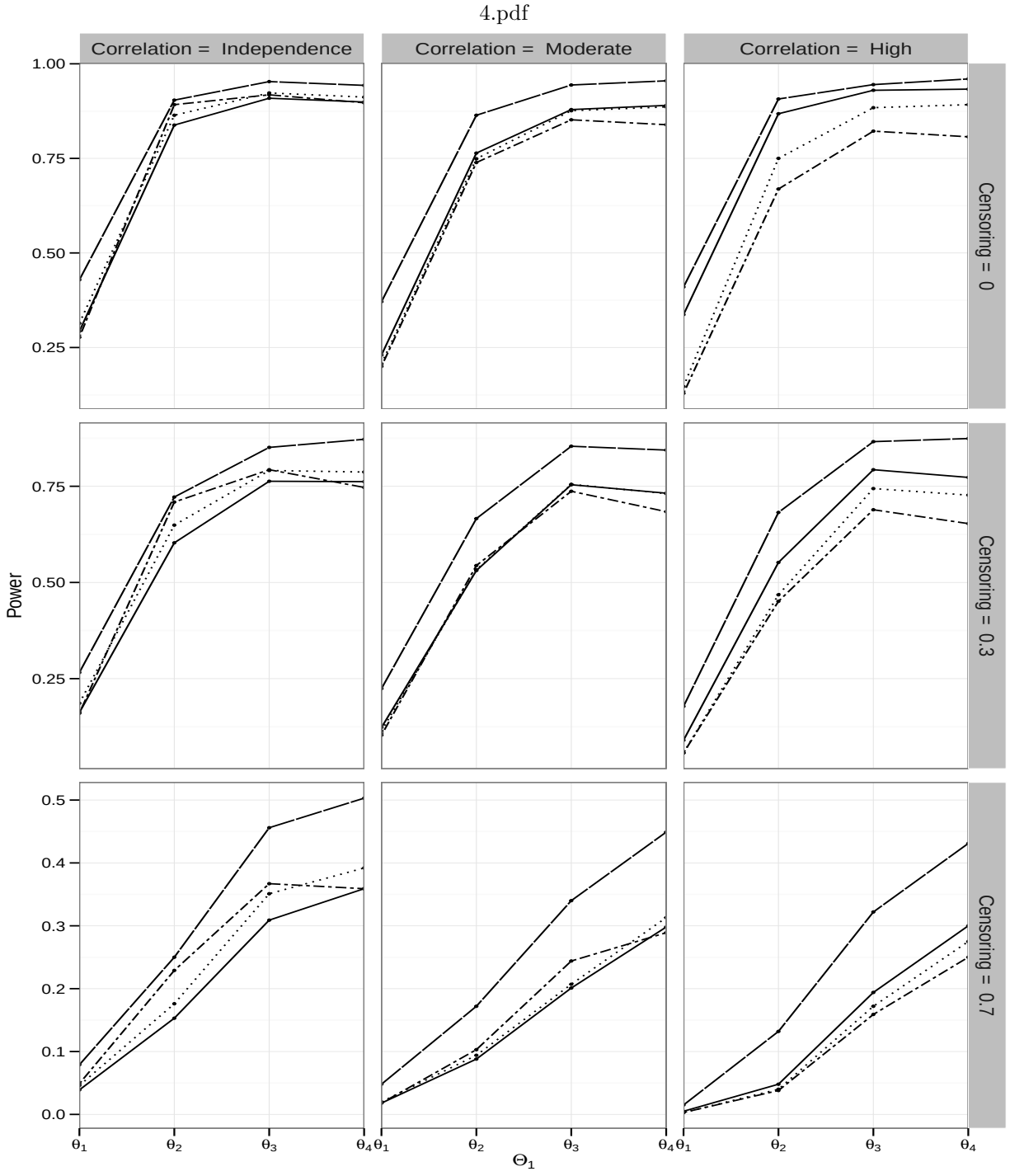

Figure 4. Power function in the case of six multivariate survival times (K = 6), $n = 800$ and $\alpha = 5\%$. Solid line (——) is the MC test, solid dotted line (– – –) is the weighted bootstrap (WB) test, doted line (·····) is the test based on confidence intervals (IBGA) and two-dashed line (-·-·-) is the likelihood ratio test (LR); see Table 1 for the values of $\boldsymbol{\theta}$'s.

5.png

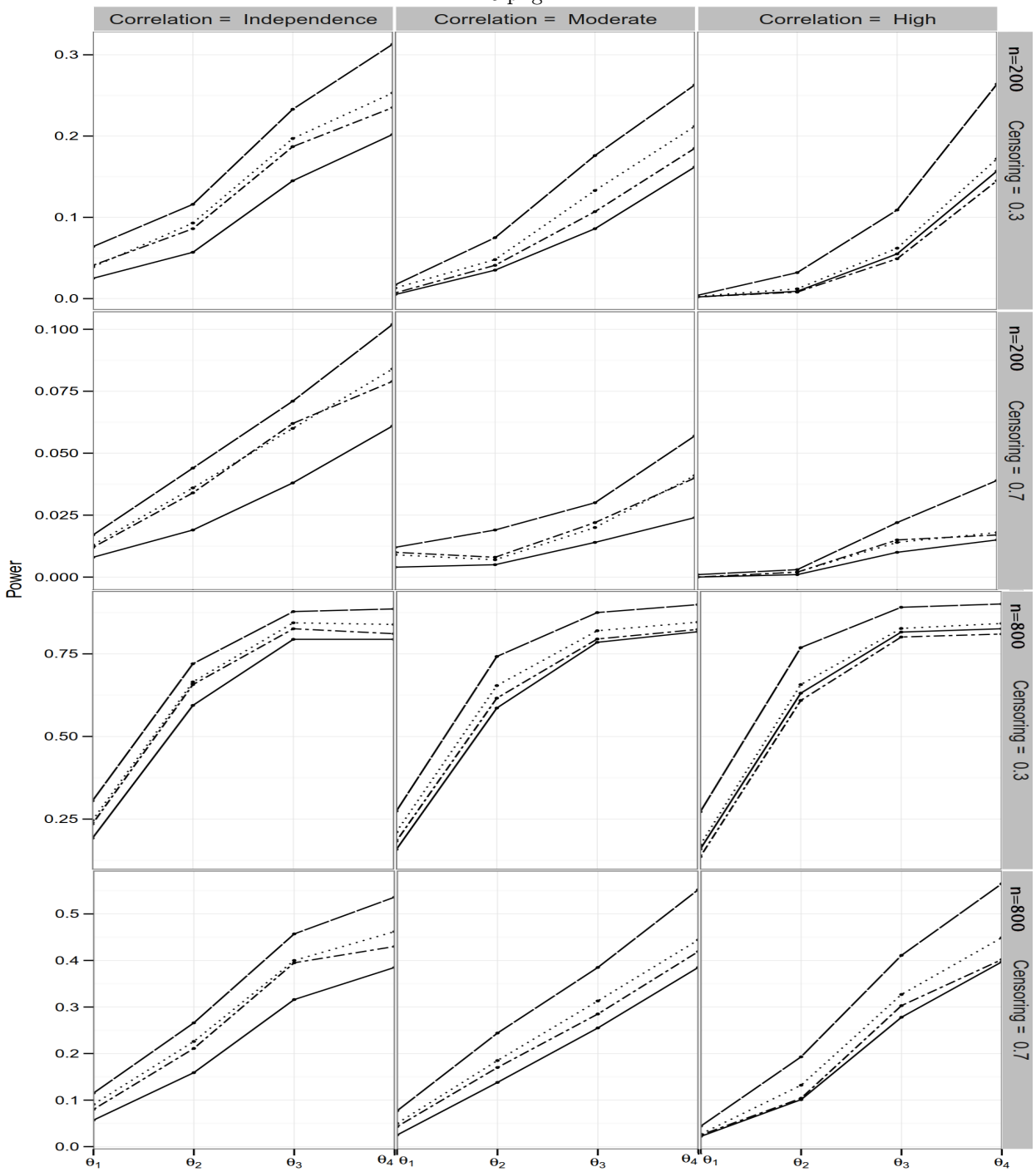

Figure 5. Power function. $K = 3$, $\alpha = 5\%$. Solid line (——) is the MC test, solid dotted line (– – –) is the weighted bootstrap (WB) test, doted line (·····) is the test based on confidence intervals (IBGA) and two-dashed line (-·-·-·) is the likelihood ratio test (LR); see Table 1 for the values of $\boldsymbol{\theta}$'s.

6.pdf

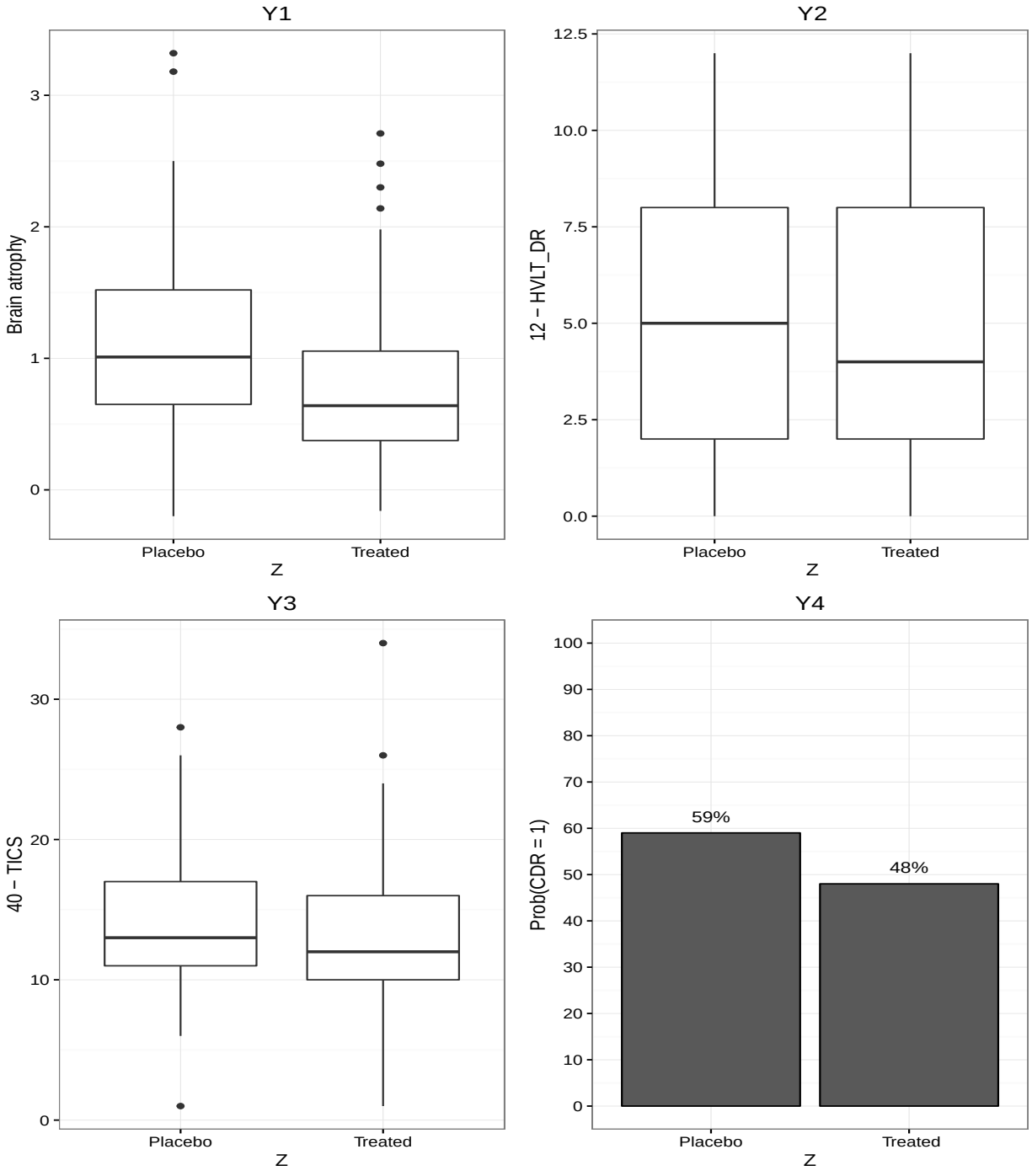

Figure 6. Descriptive plots for the Alzheimer's data.