LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Rüdiger P. Laubender and Ulrich Mansmann

Estimating individual treatment effects from responses and a predictive biomarker in a parallel group RCT

Technical Report Number 176, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Estimating individual treatment effects from responses and a predictive biomarker in a parallel group RCT

R.P. Laubender[1] and U. Mansmann[1,2]
[1]Institute of Medical Informatics, Biometry, and Epidemiology (IBE), LMU Munich
[2]Department of Statistics, LMU Munich

December 24, 2014

Contents

Abstract

When being interested in administering the best of two treatments to an individual patient i, it is necessary to know the individual treatment effects (ITEs) of the considered subjects and the correlation between the possible responses (PRs) Y_i^1 and Y_i^0 for two treatments 1 and 0. When data are generated in a parallel–group design RCT, it is not possible to determine the ITE for a single subject since we only observe two samples from the marginal distributions of these PRs and not the corresponding joint distribution due to the 'Fundamental Problem of Causal Inference' [Holland, 1986, p. 947]. In this article, we present a counterfactual approach for estimating the joint distribution of two normally distributed responses to two treatments. This joint distribution of the PRs Y_i^1 and Y_i^0 can be estimated by assuming a normal joint distribution for the PRs and by using a normally distributed baseline biomarker Z_i which is defined to be functionally related to the sum $Y_i^1 + Y_i^0$. Such a functional relationship is plausible since a biomarker Z_i and the sum $Y_i^1 + Y_i^0$ encode for the same information in a RCT, namely the variation between subjects. As a result of the interpretation of the biomarker Z_i as a proxy for the sum $Y_i^1 + Y_i^0$, the estimation of the joint distribution is subjected to some constraints. These constraints can be framed in the context of linear regressions with regard to the proportions of variances in the responses explained and with regard to the residual variation. As a consequence, a new light is thrown on the presence of treatment–biomarker interactions. We applied our approach to a classical medical data example on exercise and heart rate.

1 Introduction

Let us start with a randomized controlled trial (RCT) where two treatments denoted by 1 (for example new treatment) and 0 (for example placebo) are compared. Let Y_i^j with $j \in \{0,1\}$ be the potential response (PR) to the treatment when subject i is assigned to treatment j. However, we cannot simultaneously observe both responses Y_i^1 and Y_i^0 due to the 'Fundamental Problem of Causal Inference' [Holland, 1986, p. 947]. Thus, the observed response Y_i is given by

$$Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i) \tag{1. 1}$$

where $T_i \in \{0,1\}$ denotes the treatment allocated to subject i. Nonetheless, it is of interest to know the joint distribution of the PRs since we then could determine the 'Individual Treatment Effect' (ITE) $Y_i^1 - Y_i^0$ and thus we would know the probability a subject would benefit from either of both compared treatments. Such knowledge allows a physician to apply the best treatment for a subject. Due to the 'Fundamental Problem of Causal Inference' [Holland, 1986, p. 947], the expected difference of the PRs can only be estimated where this difference

$$\mathrm{E}[Y_i^1 - Y_i^0] = \mathrm{E}[Y_i^1] - \mathrm{E}[Y_i^0] \tag{1. 2}$$

is commonly known as 'Average Treatment Effect' (ATE) and can be estimated from the observed marginals of the PRs. If the difference $Y_i^1 - Y_i^0$ and thus the ITE is constant for every subject i then no subject–treatment additivity is present. If not, an interaction between the treatments and the subjects (subject–treatment interaction) is present.

In a RCT, it is possible to approximate the joint distribution of the PRs by using a replicated crossover design. This design is usually used for evaluating individual bioequivalence which allows separate estimates of between–subject variation, subject–treatment interactions (in this context called 'subject–formulation interaction') and within–subject variation [Senn, 2001]. Besides, unreplicated crossover trials are used to determine the correlation between the responses of two compared treatments: a positive correlation results when treatments with similar 'modes of action' are compared whereas a negative correlation results when treatments with different 'modes of actions' are compared [Cleophas, 1996a,b; Cleophas and de Vogel, 1998; Cleophas, 2000].
In situations where a crossover design is not feasible (for example eradication therapies in infectious diseases and cancer), we can only use a parallel–group RCT. When using such a design, subject effects (between–subject variation) can be approximated by biomarkers (known as 'prognostic biomarkers') and subject–treatment interactions can be approximated by treatment–biomarker interactions (known as 'predictive biomarkers') as insinuated by Senn [2001]. In drug development, often a biomarker is used for guiding treatment options. Such a biomarker is commonly labeled 'companion diagnostic (cDx)'. The most prominent example is the biomarker Her-2/neu used in breast cancer for making a decision of treating patients with trastuzumab.
Assume a trivariate normal distribution for the two PRs and the biomarker Z_i. The observation for subject i have mean vector μ_M given by

$$\begin{pmatrix} \mu_1 \\ \mu_0 \\ \mu_Z \end{pmatrix}$$

where μ_1 is the mean of Y_i^1, μ_0 is the mean of Y_i^0 and μ_Z is the mean of Z_i and variance–covariance matrix Σ_M is given by

$$\begin{pmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 & \rho_1\sigma_1\sigma_Z \\ \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 & \rho_0\sigma_0\sigma_Z \\ \rho_1\sigma_1\sigma_Z & \rho_0\sigma_0\sigma_Z & \sigma_Z^2 \end{pmatrix}$$

where σ_j^2 is the variance of Y_i^j and σ_Z^2 is the variance of Z_i. The correlation between Y_i^1 and Y_i^0 is denoted by ρ_{10}, the correlation between Y_i^j and Z_i by ρ_j. In short, the PRs and the biomarker are distributed as $(Y_i^1, Y_i^0, Z_i)' \sim \mathrm{N}[\mu_M, \Sigma_M]$. The trivariate normal distribution implies that the ATE is given by

$$\Delta = \mu_1 - \mu_0. \tag{1. 3}$$

Further, the trivariate normal distribution implies that the joint distribution of Y_i^j and Z_i is bivariate normal. Following, for example, Johnson and Wichern [1992, p. 128, 138–139], the conditional distribution of Y_i^j given Z_i is

$$Y_i^j | Z_i \sim \mathrm{N}\left[\alpha_j + \beta_j Z_i, \sigma_{j|Z}^2\right]$$

where

$$\beta_j = \rho_j \frac{\sigma_j}{\sigma_Z}, \tag{1. 4}$$

$$\alpha_j = \mu_j - \beta_j \mu_Z, \tag{1. 5}$$

$$\sigma^2_{j|Z} = \sigma^2_j (1 - \rho^2_j) \tag{1. 6}$$

hold. Since we cannot observe both PRs in a parallel–group RCT we can only estimate the quantities α_j, β_j and $\sigma^2_{j|Z}$ for both treatment groups. Thus, we can calculate the expected value of the response Y_i^j conditional on the biomarker Z_i, $\mathrm{E}[Y_i^j \mid Z_i]$, by a linear relationship described by the intercept α_j and the slope β_j. Based on the linear relationship, the expected difference $\mathrm{E}[Y_i^1 - Y_i^0 \mid Z_i]$, can be estimated. If the difference $\mathrm{E}[Y_i^1 - Y_i^0 \mid Z_i]$ is constant for every value of the biomarker Z_i then treatment–biomarker additivity will be present. If not, an interaction between the treatments and the biomarker is present. Such an interaction is commonly modeled in a parallel–group RCT by including an interaction term between treatment and biomarker in a linear regression model.
However, such an interaction term and the resulting difference $\mathrm{E}[Y_i^1 - Y_i^0 \mid Z_i]$ tells us only the half of the truth. The regression term only allows the estimation of the mean of the distribution of the ITEs $Y_i^1 - Y_i^0$ (unconditional or conditional on Z_i) but it does not allow the estimation of the variance of the ITEs since this variance depends on the correlation between the PRs. In the case of the unconditional ITEs $Y_i^1 - Y_i^0$, the variance is given by

$$\sigma^2_\Delta = \sigma^2_1 + \sigma^2_0 - 2\rho_{10}\sigma_1\sigma_0 \tag{1. 7}$$

where ρ_{10} is the sole unobservable quantity of σ^2_Δ. If no treatment–biomarker interactions or no subject–treatment interactions are present, the corresponding variances of the conditional and unconditional ITEs are zero.
In order to determine the variance of the ITE the dependence structure of the two PRs has to be known. As a consequence, it is necessary to make some assumptions about the joint distribution of the two PRs where the the dependence between the two PRs is described by a dependence measure like a correlation coefficient. In the case of normally distributed PRs, Gadbury and Iyer [2000] and Gadbury et al. [2001] rely on the trivariate normal distribution and derive bounds for the correlation coefficient ρ_{10} of the variance–covariance matrix Σ_M. The correlation coefficient ρ_{10} can be bounded due to the fact that the matrix Σ_M of the trivariate normal distribution has to be positive definite. Knowing the correlation coefficient ρ_{10} allows in consequence the bounding of the variance of the ITEs (unconditional or conditional on Z_i) as outlined by Gadbury and Iyer [2000] and Gadbury et al. [2001].
The following article studies the trivariate normal distribution as a model to evaluate the variation of ITEs when the PRs Y_i^1 and Y_i^0 and the biomarker Z_i are generated in a parallel–group RCT. But unlike Gadbury and Iyer [2000] and Gadbury et al. [2001], we define a functional relationship between the baseline biomarker Z_i and the PRs. As a consequence, a point estimator for the correlation coefficient ρ_{10} is obtainable instead of an estimate for the lower and the upper bound which usually cover a wide range of ρ_{10} and are thus less of practical value as already noticed by Lord [1955]. With a point estimate

of the correlation between the PRs, their joint distribution can be estimated and a point estimator for the variation of the ITEs can be derived. This allows the assessment of whether or not subject–treatment interactions are present. Additionally, it should also be noted that knowledge about the correlation parameter ρ_{10} can be used for planning RCTs with less subjects if this correlation is positive enabling a more rapid and economic drug development.
The article is organized as follows. In Section 2, the model for estimating the joint distribution of the PRs is presented. We assume normal marginal distributions of the responses under treatment 1 and 0, respectively. Further, assumptions on the biomarker for estimating the joint distribution are presented. In Section 3, estimators for estimating the joint distribution and corresponding variances are derived within the framework of the maximum likelihood theory. Conditions for the existence of variation in ITEs are presented. In Chapter 4, a medical data example is presented which is analyzed by the developed methodology for estimating joint distributions. In Chapter 5, the limitations of the presented methodology are discussed and an outlook for further issues of research is given.

2 A model for estimating the joint distribution

A model for estimating the joint distribution of the PRs is presented. Before starting, assumptions about the PRs with regard to the mechanism of how the data are generated are made. These assumptions lean on Cheng et al. [2009, p. 21]. Assume, firstly, that the PR of subject i is independent of the allocation of the treatments to the subjects other than subject i commonly known as 'Stable Unit Treatment Value Assumption (SUTVA)', secondly, that a subject enrolled in the RCT is an independent and identically distributed random sample from a well–defined population and thirdly, that there is independence between the allocation T_i and Z_i guaranteed by random allocation.
The following definition relates the biomarker Z_i to the ITE such that the biomarker Z_i can be used to estimate the joint distribution of Y_i^1 and Y_i^0.

Definition 2.1. The biomarker Z_i is defined as variable given by

$$Z_i = \lambda + \kappa(Y_i^1 + Y_i^0) + \eta_i \tag{2. 8}$$

where λ and κ are constants and the error term η_i is assumed to be independent of Y_i^j, $\eta_i \perp Y_i^j$, independent of $(Y_i^1 + Y_i^0)$, $\eta_i \perp (Y_i^1 + Y_i^0)$ and independent of ϵ_i^j, $\eta_i \perp \epsilon_i^j$, and identically and independently distributed (iid) as $\eta_i \mid Y_i^1 + Y_i^0 \sim \mathrm{N}[0, \sigma_\eta^2]$ where σ_η^2 denotes the variance of η_i. The distribution of the variable Z_i is as follows distributed.

Theorem 2.2. *If the random variables $(Y_i^1, Y_i^0, \eta_i)'$ have a normal distribution then the linear combination in Definition 2.1 will be normally distributed with mean value $\lambda + \kappa(\mu_1 + \mu_0)$ and variance $\kappa^2\left(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0\right) + \sigma_\eta^2$.*

Proof. The proof is standard and can be done by using characteristic functions for normally distributed random variables (see, for example, Bryc [1995, pp. 12-13]). □

Remark 2.3. In the following it is elaborated why the biomarker Z_i can be used as a proxy of the sum $Y_i^1 + Y_i^0$ and not, for example, the difference $Y_i^1 - Y_i^0$. The idea for interpreting the variable Z_i as a proxy of the sum $Y_i^1 + Y_i^0$ is best understood by a plot of the PRs Y_i^0 against Y_i^1 where a line with intercept equal to zero and slope equal to one (identity line) is added. This line indicates equality and perfect agreement between the two treatments 1 and 0 in the case that the ATE for the compared treatments 1 and 0 is zero. Following Bartko [1994, p. 741], variation alongside this line quantifies variation in the subject effects, that is 'between–subject variation' [Shumaker and Metzler, 1998, p. 1067], whereas variation orthogonal to this line quantifies variation in treatment effects among subjects which all have the same subject effect, that is 'within–subject variation' [Shumaker and Metzler, 1998, p. 1067]. Subject–treatment interactions are present when the PRs do not cluster around the line indicating equality of treatments as shown in the left plot of Figure 1.
Rotating the scatter plot graphing Y_i^0 against Y_i^1 by 45° in a clockwise manner yields a plot known as 'Tukey sum–difference graph' [Cleveland, 1985, p. 122]. This plot can be produced by graphing the differences $Y_i^1 - Y_i^0$ against the sums $Y_i^1 + Y_i^0$. The diagonal line indicating equality of treatments shown in the non–rotated scatter plot is now a line with an intercept of zero and a slope of zero. Now, variation along the x–axis and thus in the sums $Y_i^1 + Y_i^0$ represents variation in subject effects. It is well known that subject effects can be approximated by a baseline biomarker Z_i in parallel–group design RCTs (see, for example, Senn [2001]). Thus, it is reasonable to make the assumption of a functional relationship between the sum $Y_i^1 + Y_i^0$ and the biomarker Z_i since the sum $Y_i^1 + Y_i^0$ and the biomarker Z_i encode for the same information, namely the variation in the subject effects. The corresponding sum–difference graph of the scatter plot of the left plot of Figure 1 is shown in the right plot of Figure 1.

After having made the definition of the variable Z_i, the trivariate normal distribution with that variable is given as follows.

Theorem 2.4. *The joint distribution of $(Y_i^1, Y_i^0, Z_i)'$, where Z_i is interpreted as variable as defined by Definition 2.1, follows a trivariate normal distribution given by the mean vector μ_R with*

$$\begin{pmatrix} \mu_1 \\ \mu_0 \\ \lambda + \kappa(\mu_1 + \mu_0) \end{pmatrix}$$

and the variance–covariance matrix Σ_R with

$$\begin{pmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 & \kappa\sigma_1(\sigma_1 - \rho_{10}\sigma_0) \\ \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 & \kappa\sigma_0(\rho_{10}\sigma_1 - \sigma_0) \\ \kappa\sigma_1(\sigma_1 - \rho_{10}\sigma_0) & \kappa\sigma_0(\rho_{10}\sigma_1 - \sigma_0) & \sigma_Z^2 \end{pmatrix}$$

where $\sigma_Z^2 = \kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2$. This variance–covariance matrix Σ_R is non–negative definite for $\rho_{10} \in [-1, 1]$ and positive definite for $\rho_{10} \in (-1, 1)$.

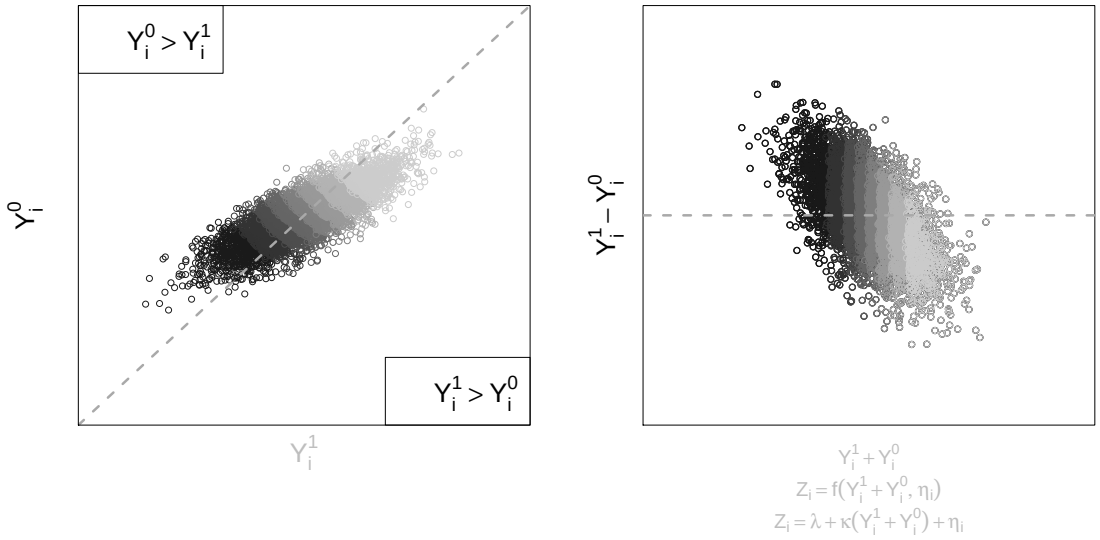

Figure 1: **Left plot:** Scatter plot of Y_i^0 on Y_i^1 with subject–treatment interactions showing between– and within–subject variation and reference line of no subject–treatment interactions (dashed grey line). **Right plot:** Corresponding sum–difference plot of the scatter plot with the assumed functional relationship of the sum $Y_i^1 + Y_i^0$ and Z_i.

Proof. The proof can be found in the Appendix. □

Now, the trivariate normal distribution is completely described by the parameters $\theta_R = (\mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$. The information about the parameter ρ_{10}, which is necessary for estimating the joint distribution of Y_i^1 and Y_i^0, is 'contained' in the correlations between Y_i^j and Z_i. These correlations can be estimated in parallel–group RCTs.

3 Maximum likelihood estimation

In the following, we propose estimators of the parameters of the trivariate normal distribution with the variable Z_i given in Definition 2.1. To this end, we represent data generated in a parallel–group RCT by (Y_i, T_i, Z_i) considering two separated treatment groups. We start with deriving maximum likelihood estimators of the parameters $(\mu_Z, \sigma_Z^2, \alpha_j, \beta_j, \sigma_{j|Z}^2)$. They will be used to derive the estimates needed for estimating the joint distribution of the PRs Y_i^1 and Y_i^0 and the variable Z_i as specified by the parameters $(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$. Further, conditions where the estimation of the joint distribution of the PRs is possible are shown. For each subject i either response Y_i^1 or Y_i^0 is observable. This results in a missing value problem with respect to the PRs Y_i^1 or Y_i^0 per subject. This was first recognized by Lord [1955] and Anderson [1957]. Their ideas are summarized in the following proposition.

Theorem 3.1. *Assume Y_i^1 and Z_i are iid as bivariate normal distribution with means μ_1 and μ_Z, variances σ_1^2 and σ_Z^2 and the correlation coefficient ρ_1. The corresponding realizations are denoted by y_i^1 and z_i and the corresponding normal densities by $\phi[z_i; \mu_Z, \sigma_Z^2] \times$*

$\phi[y_i^1;\alpha_1+\beta_1 z_i,\sigma^2_{1|Z}]$. *By analogy, assume* Y_i^0 *and* Z_i *are iid as bivariate normal distribution with means* μ_0 *and* μ_Z*, variances* σ_0^2 *and* σ_Z^2 *and the correlation coefficient* ρ_0*. The corresponding realizations are denoted by* y_i^0 *and* z_i *and the corresponding normal densities by* $\phi[z_i;\mu_Z,\sigma_Z^2]\times\phi[y_i^0;\alpha_0+\beta_0 z_i,\sigma^2_{0|Z}]$.
Let N *be the sample size and let the treatment indicator* T_i *be independently distributed from* Z_i*,* $T_i \perp Z_i$*, where the proportion* ψ *of* N *is allocated to treatment* 1 *and* $1-\psi$ *of* N *is allocated to treatment* 0*. Hence,* $\psi N = n_1$ *observations* $(y_i^1, z_i^1)'$ *and* $(1-\psi)N = n_0$ *observations* $(y_i^0, z_i^0)'$ *are given where the data are sorted by the treatment indicator leading to the following data structure (adapted from Anderson [1957, p. 202])*

$$\begin{array}{ll} z_1,\ldots,z_{n_1}, & z_{n_1+1},\ldots,z_{n_1+n_0} \\ y_1^1,\ldots,y_{n_1}^1, & \\ & y_{n_1+1}^0,\ldots,y_{n_1+n_0}^0 \end{array}$$

with $i\in\{1,\ldots,n_1,n_1+1,\ldots,n_1+n_0\}$.
The sample space $\mathcal{X}$ *of* $(Y_i^1,Z_i)'$ *and* $(Y_i^0,Z_i)'$ *is* $(\mathbf{R}^2)^{n_1}\cup(\mathbf{R}^2)^{n_0}$*. Further, the parameter space is*

$$\Theta_C \equiv \{(\mu_Z,\sigma_Z^2,\alpha_1,\alpha_0,\beta_1,\beta_0,\sigma^2_{1|Z},\sigma^2_{0|Z})(\mu_Z,\alpha_1,\alpha_0,\beta_1,\beta_0)\in\mathbf{R}^5,(\sigma_Z,\sigma_{1|Z},\sigma_{0|Z})\in\mathbf{R}^3_+\}.$$

Then the maximum likelihood estimators for μ_Z *and* σ_Z^2 *denoted by* $\hat\mu_Z$ *and* $\hat\sigma_Z^2$ *are given by*

$$\hat\mu_Z = \frac{1}{N}\sum_{i=1}^N z_i, \tag{3. 9}$$
$$\hat\sigma_Z^2 = \frac{1}{N}\sum_{i=1}^N z_i^2 - \hat\mu_Z^2. \tag{3. 10}$$

Further, the maximum likelihood estimators for β_j*,* α_j *and* $\sigma^2_{j|Z}$ *denoted by* $\hat\beta_j$*,* $\hat\alpha_j$ *and* $\hat\sigma^2_{j|Z}$ *are given by*

$$\hat\beta_j = \frac{\sum_{i=1}^{n_j}(y_i^j-\bar y_j)(z_i-\bar z_j)}{\sum_{i=1}^{n_j}(z_i-\bar z_j)^2}, \tag{3. 11}$$
$$\hat\alpha_j = \bar y_j - \hat\beta_j\bar z_j, \tag{3. 12}$$
$$\hat\sigma^2_{j|Z} = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i^j-\bar y_j)^2 - \hat\beta_j^2\frac{1}{n_j}\sum_{i=1}^{n_j}(z_i-\bar z_j)^2 \tag{3. 13}$$

where $\bar y_1 = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i^j$ *and* $\bar z_j = \frac{1}{n_j}\sum_{i=1}^{n_j} z_i$.

Proof. The proof can be found in the Appendix. □

Before proceeding to the estimation of the parameters $\theta_R = (\mu_j,\sigma_j^2,\rho_{10},\kappa,\lambda,\sigma_\eta^2)'$, we provide marginal parameters $\theta_M = (\mu_j,\mu_Z,\sigma_j^2,\sigma_Z^2,\rho_j)'$ for they are useful in the estimation of the parameters θ_R.

Theorem 3.2.

$$\mu_j = \alpha_j+\beta_j\mu_Z, \tag{3. 14}$$
$$\sigma_j^2 = \sigma^2_{j|Z}+\beta_j^2\sigma_Z^2, \tag{3. 15}$$
$$\rho_j = \frac{\beta_j\sigma_Z}{\sqrt{\sigma^2_{j|Z}+\beta_j^2\sigma_Z^2}}. \tag{3. 16}$$

Substituting the parameters of θ_M *by the corresponding estimators for* θ_C *provided in (3. 9)–(3. 13) gives the maximum likelihood estimators for* θ_M *denoted by*

$$\hat{\theta}_M = (\hat{\mu}_j, \hat{\mu}_Z, \hat{\sigma}_j^2, \hat{\sigma}_Z^2, \hat{\rho}_j)'.$$

The parameters θ_C *are transformed by the function* $f_M : \Theta_C \to \Theta_M$ *defined on the parameter space*

$$\begin{aligned}\Theta_M \equiv \{(\mu_1, \mu_0, \mu_Z, \sigma_1, \sigma_0, \sigma_Z, \rho_1, \rho_0) \mid (\mu_1, \mu_0, \mu_Z) \in \mathbf{R}^3,\\ (\sigma_1, \sigma_0, \sigma_Z) \in \mathbf{R}_+^3, (\rho_1, \rho_0) \in (-1,1)^2\}.\end{aligned}$$

Proof. The proof can be found in the Appendix. □

For estimation of the joint distribution of the PRs, the parameters $\theta_R = (\mu_j, \sigma_j^2, \rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$ have to be estimated.

Theorem 3.3. *The parameters for* μ_j *and* σ_j^2 *are already given in (3. 14)–(3. 15). The parameters* $(\rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$ *are given by*

$$\rho_{10} = \frac{\beta_1(\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2) - \beta_0(\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2)}{(\beta_0 - \beta_1)\sqrt{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\sqrt{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2}}, \tag{3. 17}$$

$$\kappa = \frac{(\beta_0 - \beta_1)\sigma_Z^2}{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 + (\beta_0^2 - \beta_1^2)\sigma_Z^2}, \tag{3. 18}$$

$$\lambda = \frac{\mu_Z(\sigma_{0|Z}^2 - \sigma_{1|Z}^2) - (\alpha_1 + \alpha_0)(\beta_0 - \beta_1)\sigma_Z^2}{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 + (\beta_0^2 - \beta_1^2)\sigma_Z^2}, \tag{3. 19}$$

$$\sigma_\eta^2 = \frac{(\sigma_{0|Z}^2 - \sigma_{1|Z}^2)\sigma_Z^2}{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 + (\beta_0^2 - \beta_1^2)\sigma_Z^2}. \tag{3. 20}$$

Substituting the parameters θ_R *by the estimators for* θ_C *provided in (3. 9)–(3. 13) gives the estimators for* θ_R *denoted by*

$$(\hat{\mu}_j, \hat{\sigma}_j^2, \hat{\rho}_{10}, \hat{\kappa}, \hat{\lambda}, \hat{\sigma}_\eta^2)'.$$

The vector θ_R *is a bijective and continuous function of* θ_C *denoted by* $f_R^C : \Theta_C^\star \to \Theta_R$ *which is defined on a subset of the parameter space* Θ_C*, in particular*

$$\begin{aligned}\Theta_C^\star \equiv \Theta_C \cap \Bigg\{ \Bigg\{ \Bigg\{ \left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2} \right) \sigma_Z^2 > 0 \Bigg\} \cap \{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 > 0\} \Bigg\}\\ \cup \Bigg\{ \Bigg\{ \left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2} \right) \sigma_Z^2 < 0 \Bigg\} \cap \{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0\} \Bigg\} \Bigg\},\end{aligned} \tag{3. 21}$$

and

$$\begin{aligned}\Theta_R \equiv \Big\{ (\mu_1, \mu_0, \sigma_1, \sigma_0, \rho_{10}, \kappa, \lambda, \sigma_\eta) \mid (\mu_1, \mu_0, \lambda) \in \mathbf{R}^3,\\ \kappa \in \mathbf{R} \backslash \{0\}, (\sigma_1, \sigma_0) \in \mathbf{R}_+^2 \backslash \{\sigma_0 - \sigma_1 = 0\},\\ \sigma_\eta \in \mathbf{R}_+, \rho_{10} \in (-1,1) \Big\}\end{aligned} \tag{3. 22}$$

When transforming the marginal parameters θ_M to the parameters θ_R by the function $f_R^M : \Theta_M^\star \to \Theta_R$ the parameter space $\Theta_M^\star$ is defined as follows:

$$\Theta_M^\star \equiv \Theta_M \cap \Big\{ \Big\{ \{\rho_0^2 - \rho_1^2 > 0\} \cap \{\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2) > 0\} \Big\} \cup \Big\{ \{\rho_0^2 - \rho_1^2 < 0\} \cap \{\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2) < 0\} \Big\} \Big\}.$$

The asymptotic variances (and covariances) of θ_R can be derived by using the multivariate delta method [Greene, 2003, Chapter D.2.7] and can be found in the Appendix.

Remark 3.4. It follows from (3. 21) and (3. 22) that the constraints

$$\left\{ \left\{ \left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2 \sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2 \sigma_Z^2} \right) \sigma_Z^2 > 0 \right\} \cap \{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 > 0\} \right\}$$

and

$$\left\{ \left\{ \left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2 \sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2 \sigma_Z^2} \right) \sigma_Z^2 < 0 \right\} \cap \{\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0\} \right\}$$

for the parameter space Θ_C and the constraints

$$\kappa \neq 0 \qquad \text{and} \qquad \sigma_0^2 - \sigma_1^2 \neq 0$$

for the parameter space Θ_R are necessary so that the function f_R^C is bijective and no discontinuity in the function f_R^C occurs and thus the estimation of the joint distribution of the PRs Y_i^1 and Y_i^0 becomes possible. As a consequence, the maximum likelihood estimate of the parameters θ_R has not to exist for some data (Y_i, Z_i, T_i).

As explained in Remark 3.4 the estimates of parameters θ_R only exist if both the difference $\hat{\rho}_0^2 - \hat{\rho}_1^2$ and the difference $\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2$ have each the same sign. Conversely, the estimates of parameters θ_R do not exist if these differences are of opposite signs.

4 Data example

A classical example of a RCT is taken from Schwenke [1990, Table 1]. The RCT randomized 24 patients to three types of exercise programs in a randomization ratio of $1:1:1$ (8 patients per exercise program). The clinically relevant endpoint is the heart rate observed after treatment where a lower value is more favorable. The biomarker Z_i is the baseline heart rate [Schwenke, 1990, p. 444]. Similar to Schwenke [1990, Figure 1], the scatter plot in Figure 2 shows the data superimposed with the linear relationships between the heart rate observed after treatment and the baseline heart rate separately for each exercise group. Following Schwenke [1990, Table 3, 'Without Bonferroni Adjustment'], the contrast exercise 1 compared to exercise 3 (contrast A) and exercise 2 compared to exercise 3 (contrast B) are considered so that in the following exercise 3 is regarded as reference treatment $(T_i = 0)$.

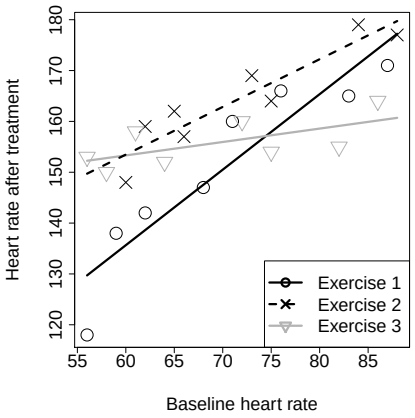

Figure 2: Plot of the heart rate observed after treatment y_i^j under exercise 1, 2, and 3 vs. baseline heart rate z_i^1 with estimated linear relationships.

Table 1: Estimates of selected parameters with corresponding asymptotic 95% confidence intervals (CIs). The lower and upper limits of the confidence intervals for some estimates are outside their range which is due to the fact that asymptotic estimators instead of exact estimators are used.

	Exercise 1 vs. 3 ($N = 16$)		**Exercise 2 vs. 3** ($N = 16$)	
	Estimate	**95% CI**	**Estimate**	**95% CI**
		Estimates of the restriction parameters:		
$\hat{\rho}_1$	0.93	0.84, 1.01	0.94	0.86, 1.01
$\hat{\rho}_0$	0.64	0.25, 1.03	0.63	0.23, 1.03
$\hat{\sigma}_{1\|Z}$	6.29	0.89, 8.85	3.58	0.51, 5.04
$\hat{\sigma}_{0\|Z}$	3.32	0.47, 4.67	3.32	0.47, 4.67
		Estimates of the parameters θ_R:		
$\hat{\mu}_1$	150.13	141.33, 158.93	163.26	157.98, 168.54
$\hat{\mu}_0$	155.88	153.21, 158.55	156.06	153.42, 158.71
$\hat{\sigma}_1$	16.82	7.93, 22.43	10.15	4.87, 13.51
$\hat{\sigma}_0$	4.32	1.08, 6.02	4.26	1.03, 5.94
$\hat{\rho}_{10}$	0.53	-0.03, 1.08	0.35	-0.32, 1.02
$\hat{\kappa}$	0.51	0.37, 0.65	0.81	0.54, 1.09
$\hat{\lambda}$	-86.29	-129.57, -43.02	-189.61	-276.69, -102.54
$\hat{\sigma}_\eta^2$	11.93	-4.31, 28.17	2.19	-17.45, 21.82
		Estimates of presentation parameters:		
$\hat{\Delta}$	-5.75	-13.74, 2.24	7.20	2.44, 11.96
$\hat{\sigma}_\Delta$	15.00	6.48, 20.19	9.55	4.52, 12.72
$\Phi\left[\hat{\Delta}/\hat{\sigma}_\Delta\right]$	0.35	0.14, 0.56	0.77	0.60, 0.95

The crossing of the both lines of exercise 1 and 3 indicates a qualitative interaction for contrast A whereas the nearly non–crossing lines of exercise 2 and 3 pronounces a quantitative interaction for contrast B. As a first step, it is checked of whether the restrictions on the parameter space $\Theta_C^\star$ are fulfilled so that the joint distribution of Y_i^1 and Y_i^0 can be estimated. It can be seen from Table 1 that for both contrasts A and B $\hat{\rho}_0^2 - \hat{\rho}_1^2 < 0$ and $\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2 < 0$ hold. Thus the estimates of the parameters θ_R will be in the corresponding parameter space Θ_R as explained by Remark 3.4.
Next, the estimates of the parameters θ_R with asymptotic 95% confidence intervals are shown in Table 1. Exercise 1 is on average lower than exercise 3 with an ATE $\hat{\Delta} = -5.75$ where the corresponding 95% confidence interval includes 0. In contrast, exercise 2 is on average higher than exercise 3 with $\hat{\Delta} = 7.20$ where the corresponding 95% confidence interval does not include 0. However, the ATE is not informative about whether or not there are patients which will more likely profit from exercise 1 (or 2) or more likely from exercise 3. In order to get this information, it is necessary to know the correlation ρ_{10} for the contrasts A and B. The correlations $\hat{\rho}_{10}$ are positive for both contrasts. The resulting variation in ITEs is described by $\hat{\sigma}_\Delta$ and we can see that the corresponding 95% confidence intervals do not include 0 for both contrasts indicating the presence of subject–treatment interactions.
Based on the estimates of the parameters θ_R, the bivariate normal density of the estimated joint distribution of the PRs Y_i^1 and Y_i^0 is drawn for both contrasts in Figure 3 (top row). From this estimated joint distribution of Y_i^1 and Y_i^0 we can quantify the subgroup of patients which will more likely benefit from exercise 1 (or 2) compared to exercise 3 and vice versa what can be seen from the location of the normal density around the identity line: In the upper wedge is the proportion of patients who will benefit from exercise 1 (or 2) whereas in the lower wedge is the proportion of patients who will benefit from exercise 3. This subgroup can be quantified by the probability that $Y_i^1 > Y_i^0$ hold, that is the probability that a higher (unfavorable) response is observed under exercise 1 (or 2) than under exercise 3. This probability is estimated by $\Phi\left[\hat{\Delta}/\hat{\sigma}_\Delta\right]$ where $\Phi[\cdot]$ denotes the cumulative distribution function of the standard normal distribution. Note that this probability heavily depends on the correlation coefficient ρ_{10}. For contrast A, $\Phi\left[\hat{\Delta}/\hat{\sigma}_\Delta\right]$ is 0.35, that is on average 35 out of 100 randomly selected patients from a population will reach a higher unfavorable response under exercise 1 than under exercise 3 and thus should be treated by exercise 3. Vice versa, on average 65 out of 100 randomly selected patients will benefit from exercise 3 compared to exercise 1. For contrast B $\Phi\left[\hat{\Delta}/\hat{\sigma}_\Delta\right] = 0.77$ is observed. A similar reasoning holds for contrast B. Although the ATE is in clear favor for exercise 3 in this case, there are nonetheless 23 out of 100 randomly selected patients which will be harmed by exercise 3.
For clinical decision–making, it is of interest to have not only estimates with Z_i 'integrated out' like $\hat{\Delta}$ and $\Phi\left[\hat{\Delta}/\hat{\sigma}_\Delta\right]$ but also to have estimates conditional on $Z_i = z_i$. Such estimates are shown in Figure 3 (bottom row) where the treatment effects conditional on Z_i are shown with 95% confidence and prediction intervals and where probabilites that $Y_i^1 > Y_i^0$ conditional on Z_i holds are shown with asymptotic 95% confidence intervals. Details on these estimators can be found in Laubender [2014]. For contrast A, it can be seen that there is a clear distinction between those patients benefitting from exercise 1 or from

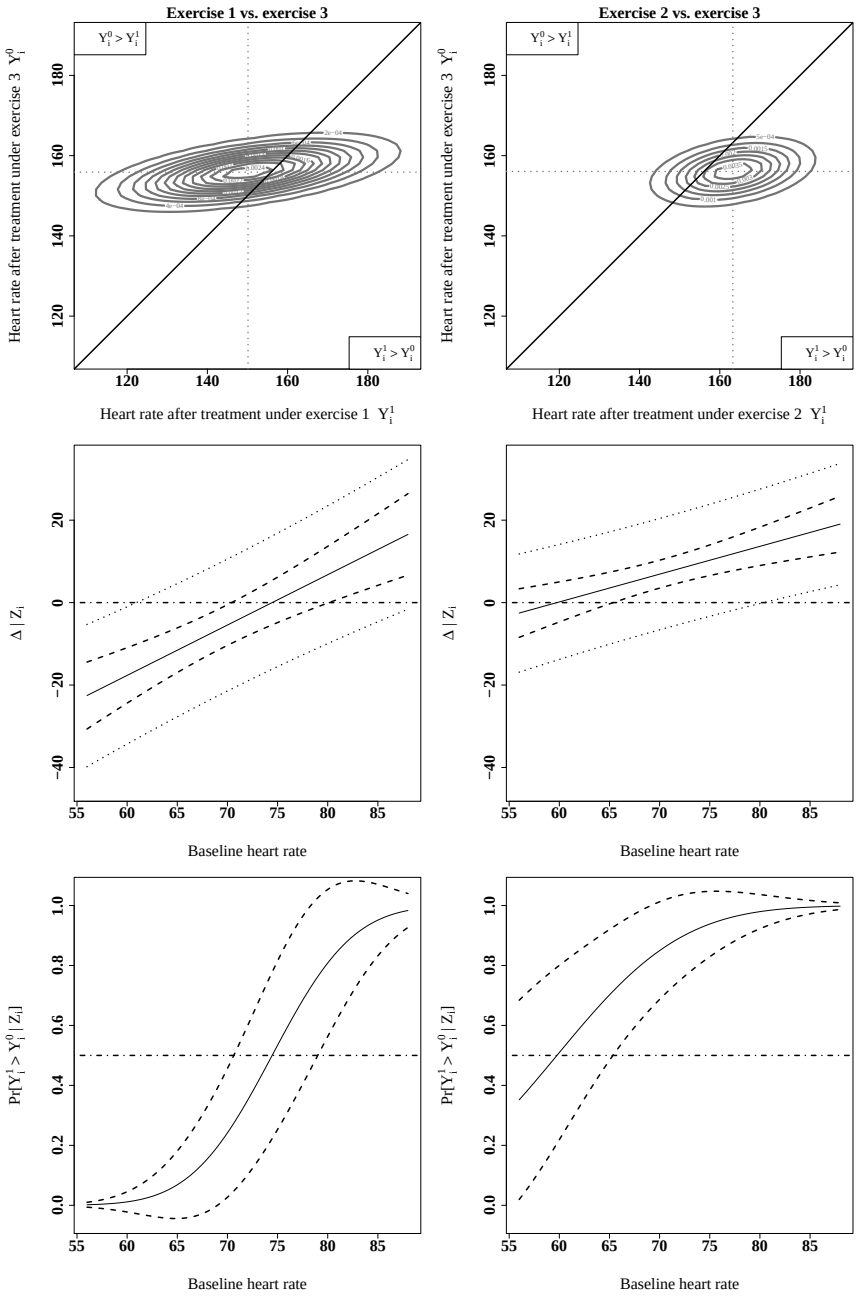

Figure 3: The left column of plots refers to contrast A whereas the right column of plots refers to contrast B. **Top row:** estimated joint distribution of heart rate after treatment Y_i^1 and Y_i^0 where the corresponding values of μ_1 and μ_0 are added by vertical and horizontal dotted lines. Further, the line of no difference between the two treatments is indicated by the diagonal; **Middle row:** plot of the ITEs conditional on the baseline heart rate Z_i (solid line) with 95% confidence intervals (dashed lines) and 95% prediction intervals (dotted lines) and with reference line (dashed–dotted line) of no treatment difference between exercises 1 and 3 and exercises 2 and 3; **Bottom row:** corresponding conditional response probabilities $P[Y_i^1 > Y_i^0 | Z_i]$ (solid line) with reference line (dashed–dotted line) of no treatment effect between exercises 1 and 3 and exercises 2 and 3.

exercise 3 illustrating the qualitative nature of the treatment–covariate interaction whereas for contrast B exercise 3 should be recommended for patients with higher values of the baseline heart rate. Also note that the width of prediction interval as essential tool for clinical decision–making depends on the correlation coefficient ρ_{10}.

5 Discussion

5.1 Conclusions

When being interested in administering the best of two treatments to an individual patient, it is necessary to know the ITEs of the considered subjects and the correlation between the PRs Y_i^1 and Y_i^0. When data are generated in a parallel–group design RCT, it is not possible to determine the ITE for a single subject since we only observe two samples from the marginal distributions of these PRs and not the corresponding joint distribution due to the 'Fundamental Problem of Causal Inference' [Holland, 1986, p. 947].
In this article, we present a counterfactual approach for estimating the joint distribution of two normally distributed responses to two treatments. This joint distribution of the PRs Y_i^1 and Y_i^0 can be estimated by assuming a normal joint distribution for the PRs and by using a normally distributed baseline biomarker Z_i which is defined to be functionally related to the sum $Y_i^1 + Y_i^0$. Such a functional relationship is plausible since a biomarker Z_i and the sum $Y_i^1 + Y_i^0$ encode for the same information in a RCT, namely the variation between subjects. Further, monotone relationships are easily transformed into linear relationships. As a result of the interpretation of the biomarker Z_i as a proxy for the sum $Y_i^1 + Y_i^0$, the estimation of the joint distribution is subjected to some constraints. These constraints should be framed in the context of linear regressions. Without loss of generality, the constraints $\hat{\rho}_0^2 - \hat{\rho}_1^2 < 0$ and $\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2 < 0$ have to be fulfilled so that maximum likelihood estimators for the parameters θ_R exist. In the context of linear regression, the quantity ρ_j^2 can be interpreted as proportion of the variance of Y_i^j explained by the biomarker Z_i and the quantity $\sigma_{j|Z}^2$ can be regarded as unexplained or residual variance.
In order to understand the constraints $\hat{\rho}_0^2 - \hat{\rho}_1^2 < 0$ and $\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2 < 0$, assume without loss of generality that we want to develop a cDx Z_i for a new treatment 1, that a high response is more favorable and that the variance σ_1^2 is higher than the variance σ_0^2 indicating the presence of subject–treatment interactions (see, for example, Cox and Reid [2000, p. 21]). In this case, it is essential to see how the comparison treatment 0 performs under that cDx Z_i so that a treatment–covariate interaction can be established. However, as outlined in the Introduction, it is not sufficient to look only to the mean of the ITEs but also to the variance of the ITEs (either unconditional or conditional on Z_i). When looking at the ITEs conditional on Z_i, three types of (qualitative) interactions can be distinguished and are exemplary shown in Figure 4. The scatter plots in Figure 4 show simulated responses of treatment 1 (gray crosses) and 0 (black circles) stratified by the cDx Z_i with corresponding regression lines superimposed. The *first* interaction is shown in the top of Figure 4 and is in accordance with the constraints $\rho_0^2 - \rho_1^2 < 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0$. For a high value of Z_i, we see that the conditional mean of treatment 1 is higher than that of treatment 0 and

that higher responses under treatment 1 than under treatment 0 can be reached.
The *second* interaction is shown in the middle of Figure 4. In this case $\rho_0^2 - \rho_1^2 > 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0$ hold so that the constraints are not fulfilled. For a high value of Z_i, we can now see that the conditional mean of treatment 0 is higher than under treatment 1 but that under treatment 1 quite a lot higher responses can be reached under treatment 1 than under treatment 0. Thus, the cDx Z_i does not capture the variation of the responses Y_i^1 as accurate as the variation of the responses Y_i^0. In this case the strict focusing on mean effects is misleading.
The *third* interaction is shown in the bottom of Figure 4. In this case $\rho_0^2 - \rho_1^2 < 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 > 0$ hold so that the constraints are not fulfilled. For very high values of Z_i, we see that the conditional mean of treatment 1 is higher than under treatment 0 and that higher responses under treatment 1 can be reached than under treatment 0. However, the lower the value of the cDx Z_i becomes, a higher value of response under treatment 0 can be reached although the mean of Y_i^1 conditional on Z_i is higher than the mean of Y_i^0 conditional on Z_i.
Besides, these constraints imply the following well–known situations of linear regression modeling where no estimation of the parameters θ_R and thus the joint distribution of the PRs Y_i^1 and Y_i^0 is possible: *First*, if $\rho_1^2 = \rho_0^2 = 0$ hold, then an uninformative biomarker Z_i is present. Nonetheless, there might be subject–treatment interactions present which cannot be modeled by the uninformative biomarker Z_i. *Second*, if $\rho_1 = \rho_0 \neq 0$ and simultaneously $\sigma_{1|Z}^2 = \sigma_{0|Z}^2$ hold then subject–treatment additivity is usually assumed in this situation. *Third*, if $\rho_1 = -\rho_0$ or $\rho_0 = -\rho_1$ and simultaneously $\sigma_{1|Z}^2 = \sigma_{0|Z}^2$ hold then subject–treatment interactions are present where, however, no subject effects are simultaneously present. This fact is also in good accordance with the statement that subject–treatment interactions 'cannot be estimated separately from variation among the units' [Cox and Reid, 2000, p. 20]. From the point of view of the linear regression model an interaction without main effect for the biomarker Z_i is present.
Our model can be considered as a further development of the idea proposed by Gadbury and Iyer [2000] which bounds the correlation ρ_{10} using a biomarker due to the requirement of positive definitness of the variance–covariance matrix in a trivariate normal distribution. In contrast, our approach provides a point estimate of ρ_{10} (with a corresponding confidence interval).
Finally, our proposed approach facilitates a more–informed assessment of a biomarker's relevance for treatment selection than does the classical approach of testing for an interaction between marker and treatment in an ordinary regression model. As Huang et al. [2012] have demonstrated, a strong interaction coefficient is important for a biomarker to have value for treatment selection but is not useful for summarizing performance because it depends on other coefficients in the risk model as well as the functional form of the model. Therefore the interaction coefficient is not directly comparable between biomarkers (and models).
To present the dependency between biomarker value and ITE we use plots of the response function $f[z_i] = \Pr[\Delta > 0 | Z_i = z_i]$. This provides useful information to individual patients who have biomarker results in hand about their expected benefit of treatment given their biomarker measure.

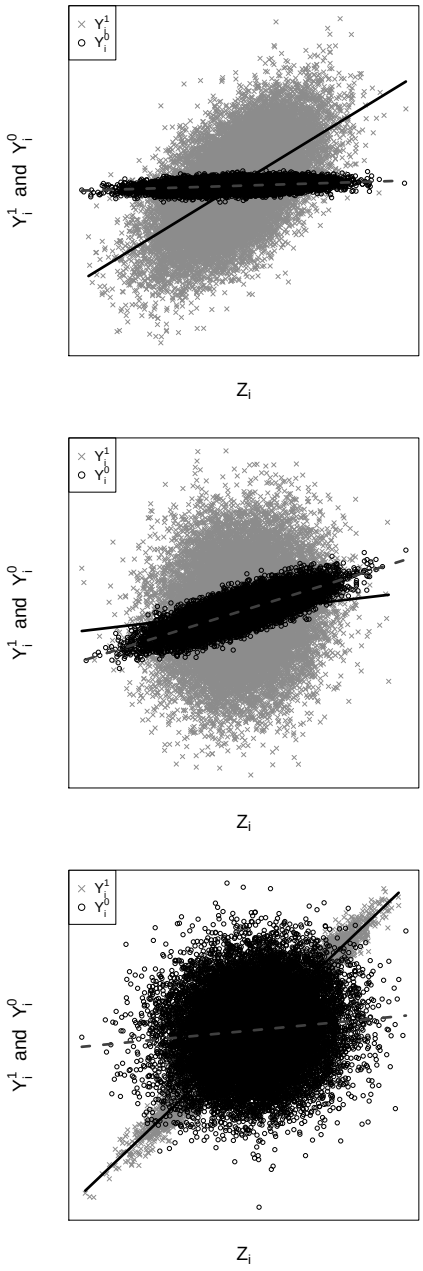

Figure 4: Scatter plots of simulated responses Y_i^j and a biomarker Z_i for treatment 1 (gray crosses) and for treatment 0 (black circles) with regression lines superimposed. All three plots show a qualitative interaction. Only the top plot shows data which fulfill the constraints $\rho_0^2 - \rho_1^2 < 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0$.

Huang et al. [2012] propose the use of a ROC curve to characterize and compare biomarkers with respect to their treatment–selection capacity. Following their idea, we can define (assuming that high biomarker values favor treatment 1) the true positive fraction $\text{TPF}[z_i] = \Pr[Z_i > z_i|\Delta > 0]$ and the false positive fraction $\text{FPF}[z_i] = \Pr[Z_i > z_i|\Delta < 0]$. The ROC curve is given by $\text{ROC}[t] = \text{TPF}[\text{FPF}^{-1}[t]]$. The TPF (as well as the FPF) can be calculated from the function $f[\cdot]$ and the distribution of the biomarker Z_i. As in the work of Huang et al. [2012], our measures provide an overview of treatment–selection capacity allowing the ITE threshold to vary. This is helpful in situations where there does not exist a well–established decision threshold and the choice relies on other factors such as the cost and side–effects of the active treatment.
Most of the methodological literature on treatment–selection markers assumes that the statistical interaction between marker value and treatment assignment in the context of a randomized trial is the primary measure of marker performance (as explained by some references of Huang et al. [2012], in particular Sargent et al. [2005], Freidlin and Simon [2005], Buyse et al. [2007], Simon [2008] and Simon et al. [2009]). However, a strong interaction is important but not sufficient for adequate marker performance [Janes, 2011]. Specifically, two markers can have the same interaction but very different performance. Huang et al. [2012] present an example where two biomarkers have the same interaction coefficient but different capacity in terms of classifying a subject according to treatment effectiveness. Therefore, we see the response function as a major information which has to be communicated to clinicians (see Figure 3).

5.2 Limitations and issues for future research

The main limitation of our approach consists in its distributional assumptions. Since simultaneous observations of both PRs Y_i^1 and Y_i^0 for a subject i are not possible it is a challenge to assess the distributional assumptions, especially that the joint distribution of the PRs follows a bivariate normal distribution. We consider copula models as alternative to the bivariate normal distribution of the PRs. They allow normal marginal distributions of the biomarker Z_i and the PRs Y_i^1 and Y_i^0 and they allow bivariate distributions of Y_i^1 and Z_i and of Y_i^0 and Z_i to be normal. Conditional copula models can be used to work out such distributions [Veraverbeke et al., 2011]. It is important to stress that these observable and assessable relationships—even made normally distributed by a transformation—do not necessarily imply that the PRs follow a joint normal distribution.
A similar critical aspect consists in the distributional assumptions of the variable Z_i created by its linear relationship with the sum $Y_i^1 + Y_i^0$. Here, it is an issue of research to explore the potential of proper scoring rules (see for example Gneiting and Raftery [2007]) to assess the correct functional relationship between biomarker Z_i and the PRs Y_i^1 and Y_i^0.
It is also of interest to study models where the correlation between the PRs Y_i^1 and Y_i^0 may depend on a biomarker value. This aspect is not studied so far and is an issue for future research. Finally, we would like to mention that the model is suited for univariate biomarkers. It is an issue for future research to generalize to the use of multiple biomarkers simultaneously. It is also of interest to apply the basic idea of this paper to other common clinical outcomes like binary endpoints and survival times, especially to remove restrictions

as introduced by Huang et al. [2012] to make the counterfactual model identifiable.

A Acknowledgements

The text of this article is based on—and in some cases plagiaristically based on—the dissertation of R.P. Laubender [Laubender, 2014] with the kind permission of the Faculty of Medicine of the LMU Munich, Germany, in order to present the results to a wider audience and with the permission of Dr. Hut Verlag, Munich.

B Appendix

B.1 Proof of Theorem 2

Proof. The mean vector (9) follows from μ_M and Theorem 1. The variances and covariances between the responses Y_i^1 and Y_i^0 shown in the variance–covariance matrix Σ_R follow from Σ_M. The variance of the variable Z_i shown in the variance–covariance matrix Σ_R follows from Proposition 1. The covariance between the response Y_i^1 and the variable Z_i is given as follows.

$$\begin{aligned}
\operatorname{Cov}[Y_i^1, Z_i] = &\quad \operatorname{Cov}[Y_i^1, \lambda + \kappa(Y_i^1 + Y_i^0) + \eta_i] \\
= &\quad \operatorname{Cov}[Y_i^1, \kappa Y_i^1 + \kappa Y_i^0] \\
= &\quad \operatorname{Cov}[Y_i^1, \kappa Y_i^1] + \operatorname{Cov}[Y_i^1, \kappa Y_i^0] \\
= &\quad \kappa\left(\operatorname{Cov}[Y_i^1, Y_i^1] + \operatorname{Cov}[Y_i^1, Y_i^0]\right) \\
= &\quad \kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0).
\end{aligned}$$

Note that the covariance between the response Y_i^1 and the error term η_i is assumed to be zero, $\operatorname{Cov}[Y_i^1, \eta_i] = 0$ as shown in Definition 1. The covariance between the response Y_i^0 and the variable Z_i can be derived by analogous line of reasoning to the derivation of the covariance between response Y_i^1 and the variable Z_i.
A matrix will be non–negative definite if the principal minors of this matrix are non–negative [Searle, 1982, pp. 205–208]. It can be seen that the principal minors of the variance–covariance matrix Σ_R are all non–negative what can be seen from

$$\begin{aligned}
\sigma_1^2 &> 0 \\
\sigma_1^2\sigma_0^2(1-\rho_{10}^2) &\geq 0 \\
\sigma_1^2\sigma_0^2\sigma_\eta^2(1-\rho_{10}^2) &\geq 0.
\end{aligned}$$

The determinant of the variance–covariance matrix Σ_R is thus given by $\sigma_1^2\sigma_0^2\sigma_\eta^2(1-\rho_{10}^2)$. If the determinant of a matrix is not zero then the matrix is nonsingular and thus invertible (see for example Searle [1982, p. 172]). It can be seen that the determinant of the variance–covariance matrix Σ_R is zero if $|\rho_{10}| = 1$. However, the density of the trivariate normal distribution requires that the variance–covariance matrix Σ_R is invertible. Thus, the variance–covariance matrix Σ_R is invertible if $\rho_{10} \in (-1, 1)$. □

B.2 Proof of Theorem 3

Proof. Denote the conditional parameters $\theta_C = (\mu_Z, \sigma_Z^2, \alpha_1, \alpha_0, \beta_1, \beta_0, \sigma_{1|Z}^2, \sigma_{0|Z}^2)'$. Consider the likelihood function $L : \Theta_C \times \mathcal{X} \to \mathbf{R}_+$ of which a value of L at θ_C is given by

$$\mathrm{L}[\theta_C] = \prod_{i=1}^{N} \phi[z_i] \prod_{i=1}^{n_1} \phi[y_i^1 \mid z_i] \prod_{i=n_1+1}^{n_1+n_0} \phi[y_i^0 \mid z_i]. \tag{B. 23}$$

The likelihood function $\mathrm{L}[\cdot]$ is splitted in three functions which have no parameters in common so that each function can be maximized independent of the other functions. Taking the natural logarithm of each of the likelihood functions and maximizing these functions for the parameters θ_C yields the estimators (9)–(13). It can be shown that these estimators form a maximum. The corresponding asymptotic variances and covariances are derived by relying on the expected information (for further details see Laubender [2014]). □

B.3 Proof of Theorem 4

Proof. The parameters $(\mu_j, \sigma_j^2, \rho_j)'$ are obtained by solving (4)–(6) for these parameters. For the parameters $(\mu_Z, \sigma_Z^2)'$ the identity function is used. It can be immediately seen that the parameters θ_M are a continuous and bijective function f_M of the parameter space Θ_C to the parameter space Θ_M. Since maximum likelihood estimators are invariant to bijective transformations [Davidson and MacKinnon, 1993, pp. 253–255], the resulting estimators of the marginal parameters θ_M are again maximum likelihood estimators. □

B.4 Proof of Theorem 5

Proof. The estimators for $(\mu_j, \sigma_j^2)'$ are obtained by solving (4)–(6). To obtain estimators for $(\rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$ in terms of the conditional parameters θ_C the following set of equations

has to be solved for these parameters:

$$\begin{aligned}\mu_Z &= \lambda + \kappa(\mu_1 + \mu_0) \\ &= \lambda + \kappa\Big((\alpha_1 + \beta_1\mu_Z) + (\alpha_0 + \beta_0\mu_Z)\Big), \qquad \text{(B. 24)}\\ \sigma_Z^2 &= \kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2 \\ &= \kappa^2\Big((\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2) + (\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2) \\ &+2\rho_{10}\sqrt{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\sqrt{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2}\Big) + \sigma_\eta^2, \qquad \text{(B. 25)}\\ \beta_1 &= \frac{\kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2} \\ &= \frac{\kappa\sqrt{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\Big(\sqrt{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2} + \rho_{10}\sqrt{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2}\Big)}{\sigma_Z^2}, \qquad \text{(B. 26)}\\ \beta_0 &= \frac{\kappa\sigma_0(\sigma_0 + \rho_{10}\sigma_1)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2} \\ &= \frac{\kappa\sqrt{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2}\Big(\sqrt{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2} + \rho_{10}\sqrt{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\Big)}{\sigma_Z^2} \qquad \text{(B. 27)}\end{aligned}$$

where the definitions of ρ_j, μ_Z and σ_Z^2 follow from Proposition 2. Solving this set of equations leads to estimators of $(\rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$ in terms of the conditional parameters θ_C. The parameters θ_R are a continuous and bijective function f_R^C of the parameter space $\Theta_C^\star$ to the parameter space Θ_R. Since maximum likelihood estimators are invariant to bijective transformations [Davidson and MacKinnon, 1993, pp. 253–255], the resulting estimators of the parameters θ_R are again maximum likelihood estimators.
To ensure that the estimators of the parameters θ_R are maximum likelihood estimators with corresponding asymptotic properties, the invariance property of maximum likelihood estimators [Davidson and MacKinnon, 1993, pp. 253–255] can be used. For that purpose, the fact that the function $f_M : \Theta_C \to \Theta_M$ is a continuous and bijective function of the conditional parameters θ_C, is at first used. Thus, when relying on the invariance principle, it has to be shown that the function $f_R^M : \Theta_M^\star \to \Theta_R$ is based both on a continuous function and that an inverse of that function exists.

The function f_R^M links the marginal parameters θ_M with the parameters θ_R as follows:

$$\begin{aligned}
\mu_1 &= \mu_1, \\
\mu_0 &= \mu_0, \\
\sigma_1^2 &= \sigma_1^2, \\
\sigma_0^2 &= \sigma_0^2, \\
\rho_{10} &= \frac{\rho_1\sigma_0 - \rho_0\sigma_1}{\rho_0\sigma_0 - \rho_1\sigma_1}, \\
\kappa &= \frac{(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{\sigma_0^2 - \sigma_1^2}, \\
\lambda &= \mu_Z - \frac{(\mu_1 + \mu_0)(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{\sigma_0^2 - \sigma_1^2}, \\
\sigma_\eta^2 &= \frac{\left(\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2)\right)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2}.
\end{aligned}$$

Lemma B.1 (Continuity of the function f_R^M)**.** *The function f_R^M is continuous if the parameter space Θ_M is restricted by either*

$$\rho_0^2 - \rho_1^2 > 0 \quad \textit{and} \quad \sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2) > 0$$

or

$$\rho_0^2 - \rho_1^2 < 0 \quad \textit{and} \quad \sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2) < 0.$$

Proof. For the parameters $(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2)'$ as functions of the parameters θ_M are identity functions which are continuous functions. The parameter ρ_{10} is a rational function and thus a continuous function (see for example Arens et al. [2012, pp. 194, 796]) provided that the denominator does not reach a value of zero. Besides, the restriction of the parameter space Θ_R for ρ_{10} has to be fulfilled. To be specific, the absolute value of the denominator has to be greater than the absolute value of the numerator. For the parameters $(\kappa, \lambda)'$ as rational functions to be continuous, it has to be ensured that no division by zero occurs. Finally, the parameter σ_η^2 is a rational function and thus a continuous function provided that the denominator does not reach a value of zero. Besides, the restriction of the parameter space Θ_R for σ_η^2 has to be fulfilled. In particular, it has to be ensured that both the numerator and the denominator are of the same signs. The restrictions on the parameter space Θ_M so that the function $f_R^M[\theta_M]$ is continuous and fulfills the requirements of the parameter space Θ_R can be derived as follows where two cases have to be distinguished:

- *Case 1: The signs of the numerator and denominator of σ_η^2 are positive.*

  In this case, the following restrictions have to be considered.

$$\begin{aligned}
\text{I}: &\quad |\rho_0\sigma_0 - \rho_1\sigma_1| > |\rho_1\sigma_0 - \rho_0\sigma_1|, \\
\text{II}: &\quad \rho_0\sigma_0 - \rho_1\sigma_1 \neq 0, \\
\text{III}: &\quad \sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2) > 0, \\
\text{IV}: &\quad \sigma_0^2 - \sigma_1^2 > 0.
\end{aligned}$$

The absolute values of inequality I are replaced by squaring. Rearranging the system of inequalities leads to

$$\begin{aligned} \text{I}: \quad & (\rho_0^2-\rho_1^2)(\sigma_0^2-\sigma_1^2)>0, \\ \text{II}: \quad & \rho_0\sigma_0-\rho_1\sigma_1\neq 0, \\ \text{III}: \quad & \sigma_0^2(1-\rho_0^2)>\sigma_1^2(1-\rho_1^2), \\ \text{IV}: \quad & \sigma_0^2-\sigma_1^2>0. \end{aligned}$$

For inequality I to be true, the inequalities $\rho_0^2>\rho_1^2$ and $\sigma_0^2>\sigma_1^2$ follow. These inequalities imply that also inequality II and inequality IV are true. Besides, inequality III imply the inequality $\sigma_0^2>\sigma_1^2$ what can be shown as follows. Since the inequality $\rho_0^2>\rho_1^2$ has to hold, ρ_0^2 can be expressed as $\rho_1^2+\xi$ where $\xi=\rho_0^2-\rho_1^2$. Thus, ξ is positive. From inequality III and the definition of ξ, it follows that

$$\frac{\sigma_0^2}{\sigma_1^2}>\underbrace{\frac{1-\rho_1^2}{1-\rho_1^2-\xi}}_{>1}. \tag{B. 28}$$

For inequality (B. 28) to be true, it can be immediately seen that $\sigma_0^2>\sigma_1^2$ has to hold. As conclusion and summary, the system of inequalities lead to the restriction of the parameter space Θ_M shown in (23) by the inequalities $\rho_0^2-\rho_1^2>0$ and $\sigma_0^2(1-\rho_0^2)-\sigma_1^2(1-\rho_1^2)>0$ so that the function f_R^M is a continuous function.

- *Case 2: The signs of the numerator and denominator of σ_η^2 are negative.*

  In this case, the following restrictions have to be considered.

$$\begin{aligned} \text{I}: \quad & |\rho_0\sigma_0-\rho_1\sigma_1|>|\rho_1\sigma_0-\rho_0\sigma_1|, \\ \text{II}: \quad & \rho_0\sigma_0-\rho_1\sigma_1\neq 0, \\ \text{III}: \quad & \sigma_0^2(1-\rho_0^2)-\sigma_1^2(1-\rho_1^2)<0, \\ \text{IV}: \quad & \sigma_0^2-\sigma_1^2<0. \end{aligned}$$

  By reasoning analogue to *Case 1*, it follows from this system of inequalities that the inequalities $\rho_0^2-\rho_1^2<0$ and $\sigma_0^2(1-\rho_0^2)-\sigma_1^2(1-\rho_1^2)<0$ have to be fulfilled for restricting the parameter space Θ_M shown in (23).

□

Lemma B.2 (Bijectivity of the function f_R^M)**.** *The function f_R^M is bijective.*

Proof. The parameters θ_R are obtained by solving $f_R^M[\theta_M]$ for θ_M leading to the function

$g_R^M : \Theta_R \to \Theta_M^\star$. The values of g_R^M at θ_R are obtained by $g_R^M[\theta_R]$:

$$\mu_1 = \mu_1, \tag{B. 29}$$
$$\mu_0 = \mu_0, \tag{B. 30}$$
$$\mu_Z = \lambda + \kappa(\mu_1 + \mu_0), \tag{B. 31}$$
$$\sigma_1^2 = \sigma_1^2, \tag{B. 32}$$
$$\sigma_0^2 = \sigma_0^2, \tag{B. 33}$$
$$\sigma_Z^2 = \kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2, \tag{B. 34}$$
$$\rho_1 = \frac{\kappa(\sigma_1 + \rho_{10}\sigma_0)}{\sqrt{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2}}, \tag{B. 35}$$
$$\rho_0 = \frac{\kappa(\sigma_0 + \rho_{10}\sigma_1)}{\sqrt{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2}}. \tag{B. 36}$$

The function f_R^M is only then an invertible function if the function f_R^M is bijective what follows from the definition of an inverse function (see for example Arens et al. [2012, pp. 36, 187]). A function is bijective if it is both injective and surjective (see for example Arens et al. [2012, pp. 36, 187]).
Injectivity is given, if $g_R^M[f_R^M[\theta_M]] = \theta_M$ for all $\theta_M \in \Theta_M^\star$ holds what is shown in the following:

$$\mu_1 = \mu_1 = \mu_1, \tag{B. 37}$$
$$\mu_0 = \mu_0 = \mu_0, \tag{B. 38}$$
$$\mu_Z = \mu_Z - (\mu_1 + \mu_0)\frac{(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{\sigma_0^2 - \sigma_1^2} + (\mu_1 + \mu_0)\frac{(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{\sigma_0^2 - \sigma_1^2} = \mu_Z, \tag{B. 39}$$
$$\sigma_1^2 = \sigma_1^2 = \sigma_1^2, \tag{B. 40}$$
$$\sigma_0^2 = \sigma_0^2 = \sigma_0^2, \tag{B. 41}$$
$$\sigma_Z^2 = \frac{(\rho_0\sigma_0 - \rho_1\sigma_1)^2\left(\sigma_1^2 + \sigma_0^2 + \frac{2\sigma_1\sigma_0(\rho_1\sigma_0 - \rho_0\sigma_1)}{\rho_0\sigma_0 - \rho_1\sigma_1}\right)\sigma_Z^2}{(\sigma_0^2 - \sigma_1^2)^2} + \frac{\left(\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2)\right)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2} = \sigma_Z^2, \tag{B. 42}$$
$$\rho_1 = \frac{\left(\sigma_1 + \frac{\sigma_0(\rho_1\sigma_0 - \rho_0\sigma_1)}{\rho_0\sigma_0 - \rho_1\sigma_1}\right)(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{(\sigma_0^2 - \sigma_1^2)\sqrt{\frac{(\rho_0^2\sigma_0^2 - \rho_1^2\sigma_1^2)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2} + \frac{\left(\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2)\right)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2}}} = \rho_1, \tag{B. 43}$$
$$\rho_0 = \frac{\left(\sigma_0 + \frac{\sigma_1(\rho_1\sigma_0 - \rho_0\sigma_1)}{\rho_0\sigma_0 - \rho_1\sigma_1}\right)(\rho_0\sigma_0 - \rho_1\sigma_1)\sigma_Z}{(\sigma_0^2 - \sigma_1^2)\sqrt{\frac{(\rho_0^2\sigma_0^2 - \rho_1^2\sigma_1^2)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2} + \frac{\left(\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2)\right)\sigma_Z^2}{\sigma_0^2 - \sigma_1^2}}} = \rho_0. \tag{B. 44}$$

Surjectivity is given, if $f_R^M[g_R^M[\theta_R]] = \theta_R$ for all $\theta_R \in \Theta_R$ holds what is shown in the following:

$$\mu_1 = \mu_1 = \mu_1, \tag{B. 45}$$
$$\mu_0 = \mu_0 = \mu_0, \tag{B. 46}$$
$$\sigma_1^2 = \sigma_1^2 = \sigma_1^2, \tag{B. 47}$$
$$\sigma_0^2 = \sigma_0^2 = \sigma_0^2, \tag{B. 48}$$
$$\rho_{10} = \frac{\rho_{10}\kappa(\sigma_0^2 - \sigma_1^2)}{\kappa(\sigma_0^2 - \sigma_1^2)} = \rho_{10}, \tag{B. 49}$$
$$\kappa = \frac{\kappa(\sigma_0^2 - \sigma_1^2)}{\sigma_0^2 - \sigma_1^2} = \kappa, \tag{B. 50}$$
$$\lambda = \lambda + \kappa(\mu_1 + \mu_0) - \frac{\kappa(\mu_1 + \mu_0)(\sigma_0^2 - \sigma_1^2)}{\sigma_0^2 - \sigma_1^2} = \lambda, \tag{B. 51}$$
$$\sigma_\eta^2 = \frac{\sigma_\eta^2(\sigma_0^2 - \sigma_1^2)}{\sigma_0^2 - \sigma_1^2} = \sigma_\eta^2. \tag{B. 52}$$

For injectivity and surjectivity to hold for the parameter spaces $\Theta_M^\star$ and Θ_R, the difference $\sigma_0^2 - \sigma_1^2$ is not allowed to be zero since a division by zero occurs in (B. 39), (B. 42)–(B. 44), (B. 49)–(B. 52). For the same reason κ and $\rho_0\sigma_0 - \rho_1\sigma_1$ are not allowed to be zero as can be seen in (B. 49) and (B. 42)–(B. 44). Further, $\rho_0^2\sigma_0^2 - \rho_1^2\sigma_1^2$ and $\sigma_0^2(1-\rho_0^2) - \sigma_1^2(1-\rho_1^2)$ are not allowed to simultaneously take a value of zero as shown by (B. 43) and (B. 44). However, it can be seen that these restrictions are already considered by the restrictions shown in Proposition B.1. Further, due to these restrictions the equation (B. 42) is positive. Therefore, the function f_R^M is bijective and the inverse function is denoted by g_R^M is the inverse of f_R^M. Thus, $\hat{\theta}_R = f_R^M[\hat{\theta}_M]$ is the maximum likelihood estimator of θ_R.

□

Lemma B.3. *The function $f_R^C[\theta_C]$ is continuous and bijective if the parameter space Θ_C is restricted by either*

$$\left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\right)\sigma_Z^2 > 0 \quad \textit{and} \quad \sigma_{0|Z}^2 - \sigma_{1|Z}^2 > 0$$

or

$$\left(\frac{\beta_0^2}{\sigma_{0|Z}^2 + \beta_0^2\sigma_Z^2} - \frac{\beta_1^2}{\sigma_{1|Z}^2 + \beta_1^2\sigma_Z^2}\right)\sigma_Z^2 < 0 \quad \textit{and} \quad \sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0.$$

Proof. The function f_M maps Θ_C to Θ_M in a continuous and bijective way. As consequence, the function also maps the corresponding subsets $\Theta_C^\star$ to $\Theta_M^\star$ in a continuous and bijective way. The function f_R^M maps the restricted parameter space $\Theta_M^\star$ to Θ_R in a continuous and bijective way. Since the composition of continuous and bijective functions leads again to continuous bijective functions (see for example Arens et al. [2012, p. 33]), the composition of the continuous and bijective functions $f_R^M \circ f_M : \Theta_C^\star \to \Theta_R$ leads to the continuous and bijective function f_R^C whose inverse is denoted by $g_R^C : \Theta_R \to \Theta_C^\star$. □

The parameters θ_C can be obtained by solving $f_R^C[\theta_C]$ for θ_C leading to the function g_R^C. For the sake of completeness, the values of g_R^C at θ_C are given in the following by $g_R^C[\theta_R]$:

$$\begin{aligned}
\mu_Z &= \lambda + \kappa(\mu_1 + \mu_0),\\
\sigma_Z^2 &= \kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2,\\
\alpha_1 &= \mu_1 + \frac{\kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0)\Big(\lambda + \kappa(\mu_1 + \mu_0)\Big)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2},\\
\alpha_0 &= \mu_0 + \frac{\kappa\sigma_0(\sigma_0 + \rho_{10}\sigma_1)\Big(\lambda + \kappa(\mu_1 + \mu_0)\Big)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2},\\
\beta_1 &= \frac{\kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2},\\
\beta_0 &= \frac{\kappa\sigma_0(\sigma_0 + \rho_{10}\sigma_1)}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2},\\
\sigma_{1|Z}^2 &= \sigma_1^2\left(1 - \frac{\kappa^2(\sigma_1 + \rho_{10}\sigma_0)^2}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2}\right),\\
\sigma_{0|Z}^2 &= \sigma_0^2\left(1 - \frac{\kappa^2(\sigma_0 + \rho_{10}\sigma_1)^2}{\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2}\right).
\end{aligned}$$

□

B.5 Asymptotic variances of the estimators of the parameters θ_R

The asymptotic variance of $\hat{\mu}_1$ can be estimated by

$$\mathrm{Var}[\hat{\mu}_1] = \frac{\hat{\sigma}_{1|Z}^2}{n} + \frac{\hat{\beta}_1^2\hat{\sigma}_Z^2}{N}.$$

The asymptotic variance of $\hat{\mu}_0$ can be estimated by

$$\mathrm{Var}[\hat{\mu}_0] = \frac{\hat{\sigma}_{0|Z}^2}{m} + \frac{\hat{\beta}_0^2\hat{\sigma}_Z^2}{N}.$$

The asymptotic variance of $\hat{\sigma}_1^2$ can be estimated by

$$\mathrm{Var}[\hat{\sigma}_1^2] = 2\left(\frac{\hat{\sigma}_{1|Z}^2(\hat{\sigma}_{1|Z}^2 + 2\hat{\beta}_1^2\hat{\sigma}_Z^2)}{n} + \frac{\hat{\beta}_1^4\hat{\sigma}_Z^4}{N}\right).$$

The asymptotic variance of $\hat{\sigma}_0^2$ can be estimated by

$$\mathrm{Var}[\hat{\sigma}_0^2] = 2\left(\frac{\hat{\sigma}_{0|Z}^2(\hat{\sigma}_{0|Z}^2 + 2\hat{\beta}_0^2\hat{\sigma}_Z^2)}{m} + \frac{\hat{\beta}_0^4\hat{\sigma}_Z^4}{N}\right).$$

The asymptotic variance of $\hat{\rho}_{10}$ can be estimated by

$$\begin{aligned}\mathrm{Var}[\hat{\rho}_{10}] = \frac{1}{2(\hat{\beta}_0-\hat{\beta}_1)^4(\hat{\sigma}^2_{1|Z}+\hat{\beta}_1^2\hat{\sigma}^2_Z)^3(\hat{\sigma}^2_{0|Z}+\hat{\beta}_0^2\hat{\sigma}^2_Z)^3}\Bigg\{\Big(\hat{\beta}_0(\hat{\sigma}^2_{1|Z}+\hat{\beta}_1^2\hat{\sigma}^2_Z)+\hat{\beta}_1(\hat{\sigma}^2_{0|Z}+\hat{\beta}_0^2\hat{\sigma}^2_Z)\Big)^2 \\ \times(\hat{\beta}_0-\hat{\beta}_1)^2\left(\frac{\hat{\sigma}^4_{1|Z}(\hat{\sigma}^2_{0|Z}+\hat{\beta}_0^2\hat{\sigma}^2_Z)^2}{n}+\frac{\hat{\sigma}^4_{0|Z}(\hat{\sigma}^2_{1|Z}+\hat{\beta}_1^2\hat{\sigma}^2_Z)^2}{m}+\frac{\hat{\sigma}^4_Z(\hat{\beta}_1^2\hat{\sigma}^2_{0|Z}-\hat{\beta}_0^2\hat{\sigma}^2_{1|Z})^2}{N}\right) \\ +\frac{2}{\hat{\sigma}^2_Z}\left(\frac{1}{n}\left(\hat{\sigma}^2_{1|Z}(\hat{\sigma}^2_{0|Z}+\hat{\beta}_0^2\hat{\sigma}^2_Z)^2\Big(\hat{\sigma}^2_Z\Big(\hat{\beta}_0^2\hat{\sigma}^2_{1|Z}(\hat{\beta}_0-\hat{\beta}_1)+\hat{\beta}_1^3\hat{\sigma}^2_{0|Z}\Big)-\hat{\beta}_0\hat{\sigma}^2_{1|Z}(\hat{\sigma}^2_{1|Z}-\hat{\sigma}^2_{0|Z}+\hat{\beta}_1^2\hat{\sigma}^2_Z)\Big)^2\right)\right. \\ \left.+\frac{1}{m}\left(\hat{\sigma}^2_{0|Z}(\hat{\sigma}^2_{1|Z}+\hat{\beta}_1^2\hat{\sigma}^2_Z)^2\Big(\hat{\sigma}^2_Z\Big(\hat{\beta}_1^2\hat{\sigma}^2_{0|Z}(\hat{\beta}_0-\hat{\beta}_1)-\hat{\beta}_0^3\hat{\sigma}^2_{1|Z}\Big)+\hat{\beta}_1\hat{\sigma}^2_{0|Z}(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+\hat{\beta}_0^2\hat{\sigma}^2_Z)\Big)^2\right)\right)\Bigg\}.\end{aligned}$$

The asymptotic variance of $\hat{\kappa}$ can be estimated by

$$\begin{aligned}\mathrm{Var}[\hat{\kappa}] = \frac{\hat{\sigma}^2_Z}{\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0^2-\hat{\beta}_1^2)\hat{\sigma}^2_Z\Big)^4}\Bigg\{(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})^2\left(\frac{\hat{\sigma}^2_{1|Z}}{n}+\frac{\hat{\sigma}^2_{0|Z}}{m}\right) \\ +(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\left\{2\left(\frac{(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})^2}{N}+\hat{\sigma}^2_{1|Z}\hat{\sigma}^2_{0|Z}\left(\frac{1}{n}+\frac{1}{m}\right)\right)+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\left(\frac{\hat{\sigma}^2_{1|Z}}{n}+\frac{\hat{\sigma}^2_{0|Z}}{m}\right)\right\}\Bigg\}.\end{aligned}$$

The asymptotic variance of $\hat{\lambda}$ can be estimated by

$$\begin{aligned}\mathrm{Var}[\hat{\lambda}] = \frac{\hat{\sigma}^2_Z}{\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0^2-\hat{\beta}_1^2)\hat{\sigma}^2_Z\Big)^2}\Bigg\{\frac{(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})^2}{N}+\frac{\alpha_1+\alpha_0+(\hat{\beta}_1+\hat{\beta}_0)\hat{\mu}_Z}{\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0^2-\hat{\beta}_1^2)\hat{\sigma}^2_Z\Big)^2} \\ \times\left(2(\hat{\beta}_0-\hat{\beta}_1)^2\Big(\alpha_1+\alpha_0+(\hat{\beta}_1+\hat{\beta}_0)\hat{\mu}_Z\Big)\hat{\sigma}^2_Z\left(\frac{\hat{\sigma}^4_{1|Z}}{n}+\frac{\hat{\sigma}^4_{0|Z}}{m}+\frac{(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})^2}{N}\right)\right. \\ +\frac{1}{n}\left(\hat{\sigma}^2_{1|Z}\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)\Big(2\hat{\beta}_1\hat{\mu}_Z(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})+\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)\right. \\ \left.\times(\alpha_1+\alpha_0)\Big)\right)+\frac{1}{m}\left(\hat{\sigma}^2_{0|Z}\Big(\hat{\sigma}^2_{1|Z}-\hat{\sigma}^2_{0|Z}+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)\Big(2\hat{\beta}_0\hat{\mu}_Z(\hat{\sigma}^2_{1|Z}-\hat{\sigma}^2_{0|Z})+\right. \\ \left.\left.\times\Big(\hat{\sigma}^2_{1|Z}-\hat{\sigma}^2_{0|Z}+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)(\alpha_1+\alpha_0)\Big)\right)\right)-(\hat{\beta}_0-\hat{\beta}_1)\Bigg\{(\hat{\beta}_1-\hat{\beta}_0)(\hat{\mu}_Z^2+\hat{\sigma}^2_Z)\left(\frac{\hat{\sigma}^2_{1|Z}}{n}+\frac{\hat{\sigma}^2_{0|Z}}{m}\right) \\ +\frac{\hat{\mu}_Z}{\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0^2-\hat{\beta}_1^2)\hat{\sigma}^2_Z\Big)}\left(\frac{1}{m}\left(\hat{\sigma}^2_{0|Z}\Big(2\hat{\beta}_0\hat{\mu}_Z(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})+\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}-(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)\right.\right.\right. \\ \left.\left.\left.\times(\alpha_1+\alpha_0)\Big)\right)-\frac{1}{n}\left(\hat{\sigma}^2_{1|Z}\Big(2\hat{\beta}_1\hat{\mu}_Z(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z})+\Big(\hat{\sigma}^2_{0|Z}-\hat{\sigma}^2_{1|Z}+(\hat{\beta}_0-\hat{\beta}_1)^2\hat{\sigma}^2_Z\Big)(\alpha_1+\alpha_0)\Big)\right)\right)\Bigg\}\Bigg\}.\end{aligned}$$

The asymptotic variance for $\hat{\sigma}_\eta^2$ can be estimated by

$$\mathrm{Var}[\hat{\sigma}_\eta^2] = \frac{2\hat{\sigma}_Z^4}{\left(\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2 + (\hat{\beta}_0^2 - \hat{\beta}_1^2)\hat{\sigma}_Z^2\right)^4} \left\{ \frac{(\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2)^4}{N} + \hat{\sigma}_Z^2 \left(2(\hat{\sigma}_{0|Z}^2 - \hat{\sigma}_{1|Z}^2)^2 \times \left(\frac{\hat{\beta}_1^2 \hat{\sigma}_{1|Z}^2}{n} + \frac{\hat{\beta}_0^2 \hat{\sigma}_{0|Z}^2}{m} \right) + (\hat{\beta}_0^2 - \hat{\beta}_1^2)^2 \hat{\sigma}_Z^2 \left(\frac{\hat{\sigma}_{1|Z}^4}{n} + \frac{\hat{\sigma}_{0|Z}^4}{m} \right) \right) \right\}.$$

References

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association* **52**, 200–203.

Arens, T., Hettlich, F., Karpfinger, C., Kockelkorn, U., Lichtenegger, K. and Stachel, H. (2012). *Mathematik*. Heidelberg: Spektrum Akademischer Verlag.

Bartko, J.J. (1994). General methodology II measures of agreement: a single procedure. *Statistics in Medicine* **13**, 737–745.

Bryc, W. (1995). *The Normal Distribution: Characterizations with Applications*. Heidelberg: Springer.

Buyse, M. (2007). Towards validation of statistically reliable biomarkers. *European Journal of Cancer* **Supplements 5**, 89–95.

Cheng, J., Small, D.S., Tan, Z. and Ten Have, T.R. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika* **96**, 19–36.

Cleophas, T.J.M. (1996a). Crossover trials are only useful when there is a positive correlation between the response to different treatment modalities. *British Journal of Clinical Pharmacology* **41**, 235–239.

Cleophas, T.J.M. (1996b). Criticism of cardiovascular studies with negative results due to a negative correlation. *Angiology* **47**, 139–147.

Cleophas, T.J.M. and de Vogel, E.M. (1998). Crossover studies are a better format for comparing equivalent treatments than parallel–group studies. *Pharmacy World and Science* **20**, 113–117.

Cleophas, T.J. (2000). Crossover trials should not be used to test one treatment against another treatment with a totally different chemical class/mode of action. *Journal of Clinical Pharmacology* **40**, 1503–1508.

Cleveland, W.S. (1985). *The Elements of Graphing Data*. Monterey: Wadsworth Advanced Books and Software.

Cox, D.R. and Reid, N. (2000). *The Theory of the Design of Experiments*. Boca Raton: Chapman & Hall/CRC.

Davidson, R. and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11**, 7872–7878.

Gadbury, G.L. and Iyer, H.K. (2000). Unit–treatment interaction and its practical consequences. *Biometrics* **56**, 882–885.

Gadbury, G.L., Iyer, H.K. and Allison, D.B. (2001). Evaluating subject–treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics* **11**, 313–333.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Greene, W.H. (2003) *Econometric Analysis*. Upper Saddle River: Prentice Hall.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81,** 945–960.

Huang, Y., Gilbert, P.B. and Janes, H. (2012). Assessing treatment–selection markers using a potential outcomes framework. *Biometrics* **68**, 687–696.

Janes, H., Pepe, M.S., Bossuyt, P.B. and Barlow, W.E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* **154**, 253–259.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. Prentice–Hall International: London.

Laubender, R.P. (2014). *Estimation of a joint distribution with two normally distributed treatment responses as marginals generated in a randomized controlled trial based on the parallel–group design by using a normally distributed covariate*. Dr. Hut Verlag: Munich.

Lord, F.M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association* **50**, 870–876.

Sargent, D.J., Conley B.A., Allegra C. and Collette L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* **23**, 2020–2027.

Schwenke, J.R. (1990). On the equivalence of the Johnson–Neyman technique and Fieller's Theorem. *Biometrical Journal* **32,** 441–447.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. Hoboken: Wiley.

Senn, S. (2001). Individual therapy: new dawn or false dawn? *Drug Information Journal* **35,** 1479–1494.

Shumaker, R.C. and Metzler, C.M. (1998). The phenytoin trial is a case study of 'individual bioequivalence'. *Drug Information Journal* **32**, 1063–1072.

Simon, R. (2008). The use of genomics in clinical trial design. *Clinical Cancer Research* **14**, 5954–5958.

Simon, R.M., Paik, S. and Hayes, D.F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute* **21**, 1446–1452.

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