Received: 30 November 2018 | Revised: 10 December 2019 | Accepted: 12 January 2020

DOI: 10.1002/bimj.201800370

Biometrical Journal

RESEARCH PAPER

Estimating the distribution of heterogeneous treatment effects from treatment responses and from a predictive biomarker in a parallel-group RCT: A structural model approach

Ruediger P. Laubender | **Ulrich Mansmann** | **Michael Lauseker**

Faculty of Medicine, Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Munich, Germany

Correspondence
Ulrich Mansmann, Faculty of Medicine, Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Marchioninistrasse 15, 81377 Munich, Germany.
Email: mansmann@ibe.med.uni-muenchen.de

Abstract

When the objective is to administer the best of two treatments to an individual, it is necessary to know his or her individual treatment effects (ITEs) and the correlation between the potential responses (PRs) Y_i^1 and Y_i^0 under treatments 1 and 0. Data that are generated in a parallel-group design RCT does not allow the ITE to be determined because only two samples from the marginal distributions of these PRs are observed and not the corresponding joint distribution. This is due to the "fundamental problem of causal inference." Here, 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 bivariate normal distribution for the PRs and by using a normally distributed baseline biomarker Z_i 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 an RCT, namely the variation between subjects. The estimation of the joint trivariate 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. This presents new insights on the presence of treatment–biomarker interactions. We applied our approach to example data on exercise and heart rate and extended the approach to survival data.

KEYWORDS

average treatment effect, individual treatment effect, reconstruction variable, subject–treatment interaction, treatment–biomarker interaction

1 | INTRODUCTION

Numerous examples in the medical literature demonstrate the existence of different treatments for the same disease with similar or different modes of action. While in some indications, one treatment is uniformly superior to the other, there are also cases, where a percentage of subjects benefited more from the one specific treatment compared to alternatives. This is called a heterogeneous

treatment effect or subject–treatment interaction. The identification of these effects within respective subjects allows selecting the appropriate individual treatment. These effects play a key role in the concept of personalized medicine.

Starting with a randomized controlled trial (RCT) that allocates two treatments denoted by 1 (for the new treatment) and 0 (for the standard treatment or placebo), let Y_i^j be the potential response (PR) to the treatment j of subject i (with $j \in \{0, 1\}$). However, we cannot simultaneously observe both responses Y_i^1 and Y_i^0 on subject i due to the "fundamental problem of causal inference" (Holland, 1986, p. 947). Formally, the observed response Y_i is given by

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

where $T_i \in \{0; 1\}$ denotes the treatment allocated to subject i. This equation holds true due to the consistency assumption, which binds the potential to the observed response.

Nonetheless, it is of interest to know the joint distribution of the PRs from which the "individual treatment effect" (ITE) $Y_i^1 - Y_i^0$ is derived and to determine the subject's probability to benefit from which treatment best. This is a decision aid for a physician to make the best treatment choice within two (or several) treatment options. But due to the "fundamental problem of causal inference" (Holland, 1986, p. 947), only the average of the expected difference of the PRs can be estimated when no further information is available. The term

$$E[Y^1 - Y^0] = E[Y^1] - E[Y^0] = \tau_1 - \tau_0$$

is commonly known as "average treatment effect" (ATE) and can be estimated from the observed marginals of the PRs. If the ITE, $Y_i^1 - Y_i^0$, is not constant within the trial population, an interaction between treatments and the subjects (subject–treatment interaction) is present.

It is possible to approximate the joint distribution of the PRs in a replicated randomized crossover design. This design is commonly used for evaluating individual bioequivalence. It allows separate estimates of between–subject variation, subject–treatment interactions (in the bioequivalence context called "subject–formulation interaction") and within–subject variation (Senn, 2001). Unreplicated crossover trials can also be used to determine the correlation between responses under two treatments. A positive correlation reflects treatments with similar "modes of action," a negative correlation reflects treatments with different "modes of actions" (Cleophas, 2000).

In situations where a crossover design is not feasible (e.g., eradication therapies in infectious diseases and cancer), a parallel–group RCT remains the only option to compare treatment effects. Here, 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 proposed by Senn (2001). Predictive biomarker guide treatment options and are labeled "companion diagnostic (cDx)." The most prominent example is the biomarker Her-2/neu. If a breast cancer patient is treated with trastuzumab depends on the value measured (Hudis, 2007).

A treatment–biomarker interaction allows assessing if a biomarker is predictive or not. Here, two kinds of interactions are distinguished: *quantitativ*e and *qualitative* interaction. A *quantitative* interaction is observed if the treatment effects differ with respect to their size but not with respect to their sign over the range of biomarker values. A *qualitative* interaction is observed if the treatment effects differ with respect to their size as well as their sign over the range of biomarker values. From a treatment point of view, only *qualitative* interactions are of interest since there are subgroups of patients who benefit from treatment 1 as well as subgroups of patients who benefit from treatment 0. In case of a *quantitative* interaction, all patients will benefit from the same treatment but in different extent.

These effects can be seen by looking at the conditional treatment effect. It is the average treatment difference given a specific biomarker value. In general, it is estimated by the difference of two regression functions which fit the biomarker value to the specific treatment response. The goal of this paper is to sharpen this concept by introducing the correlation between the two individual potential responses in order to predict the treatment effect for an individual with a given biomarker value. For this aim, we derive prediction intervals. The consequence of not considering the dependence between the responses Y_i^1 and Y_i^0 is shown in the bottom left plot of Figure 1. In this plot, the prediction intervals for a positive correlation between Y_i^1 and Y_i^0 ($\rho_{10} = 0.8$) (dashed lines) are much closer around the conditional treatment effects compared to the prediction intervals where independence (in this case thus a correlation $\rho_{10} = 0.0$) is assumed (dotted lines). The prediction intervals are wider for a negative correlation ($\rho_{10} = -0.5$) (dashed–dotted lines) than those based on a zero correlation and much wider than those based on a positive correlation.

Homogeneous treatment effects are present for $\rho_{10} = 1$. As a consequence, the treatment effect is additive and is constant in the trial population. *Heterogeneous treatment effects* are present if $\rho_{10} \neq 1$. Here, treatment effects differ among the patients. Knowing the joint distribution of the responses Y_i^1 and Y_i^0 formalizes the aims of "personalized medicine": Applying the best

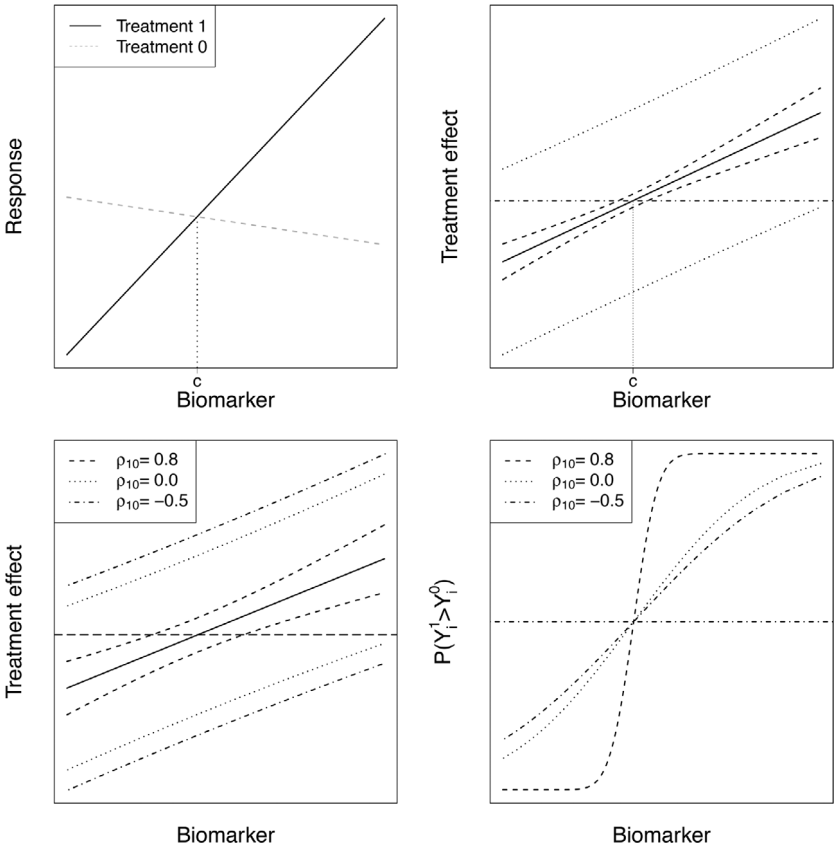

FIGURE 1 In the top left, the outcome for two different treatments is depicted, c is the value of the biomarker where both treatments are equally effective. The top right figure shows the ATE conditional on the biomarker together with the confidence and the prediction interval under the assumption of $\rho_{10} = 0$. The bottom left figure shows prediction intervals for different values of ρ_{10}. The bottom right figure shows the probailities $P(Y_i^1 > Y_i^0)$ for different values of ρ_{10}

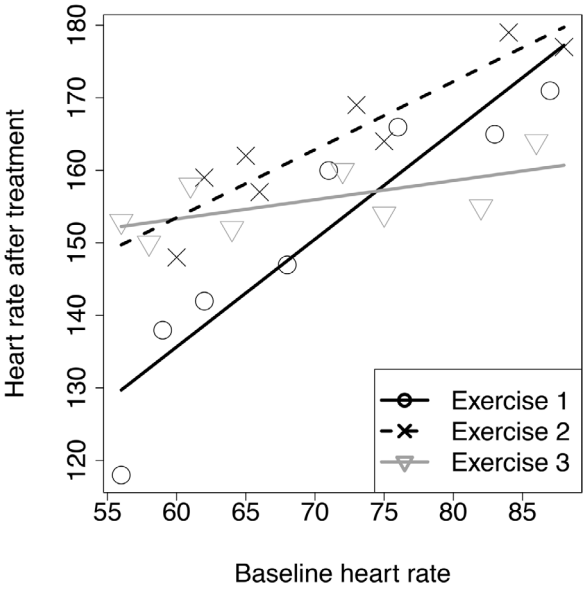

FIGURE 2 Plot of the heart rate observed after treatment y_i^j under exercise 1, 2, and 3 versus baseline heart rate z_i^1 with estimated linear relationships

of two (or more) treatment alternatives to a patient i based on his or her baseline biomarker value $Z_i = z_i$. We assume a linear dependence between the potential responses Y_i^j and the biomarker.

An example of Schwenke (1990, Table 1) demonstrates our ideas. The RCT randomized 24 patients equally to three different exercise programs to improve the heart rate. After 8 weeks of training, each participant's heart rate was recorded after a 6 min run. Schwenke also provides data on the resting heart rate at baseline, which we consider as the biomarker of interest. Lower heart rates after the exercise are considered as more favorable. Figure 2 shows the data. It indicates a linear relationship between biomarker and outcome as well as a quantitative and qualitative interaction. This example is discussed in more detail in Section 4.

The article is organized as follows. In Section 2, the model for reconstructing the joint distribution of the PRs is presented. We assume normal marginal distributions of the responses under treatment 1 and 0, respectively. Further, the assumptions on the biomarker for reconstructing the joint distribution are presented. In Section 3, estimators for reconstructing the joint distribution and corresponding variability are derived using the maximum likelihood theory. Conditions for the existence of variation in ITEs are presented. In Section 4, a medical data example is presented which is analyzed by the developed methodology for reconstructing joint distributions. In Section 5, the limitations of the presented methodology are discussed and an outlook for further issues of research is given.

2 | A MODEL FOR RECONSTRUCTING THE JOINT DISTRIBUTION

2.1 | General setting

In accordance with Cox and Reid (2000, p. 20), we extend the *potential responses* as follows:

$$\begin{aligned} Y_i^1 &= \tau_1 + U_i^1 \\ Y_i^0 &= \tau_0 + U_i^0 \end{aligned} \tag{2}$$

with τ_j as constant mean parameters and $(U_i^1, U_i^0)'$ as independent and identically distributed from a bivariate normal distribution with $E(U_i^j) = 0$. We denote the potential outcome of individual i dependent on treatment a, $a \in 0, 1$ by Y_i^a. Substituting Y_i^1 and Y_i^0 in (1) yields

$$Y_i^a = aY_i^1 + (1-a)Y_i^0 = \tau_0 + (\tau_1 - \tau_0)a + U_i^0 + \left(U_i^1 - U_i^0\right)a, \tag{3}$$

where τ_0 can be interpreted as global mean, $(\tau_1 - \tau_0)a$ as treatment effect, U_i^0 as subject effect and $(U_i^1 - U_i^0)a$ as subject–treatment interaction. This can be interpreted as a structural model of heterogeneous treatment effects considering subject-specific effects.

In an RCT based on the parallel-group design, we can approximate the potential responses Y_i^1 and Y_i^0 in (2) by a covariate Z_i and a linear relationship leading to

$$\begin{aligned} Y_i^1 &= \alpha_1 + \beta_1 Z_i + \epsilon_i^1 \\ Y_i^0 &= \alpha_0 + \beta_0 Z_i + \epsilon_i^0 \end{aligned} \tag{4}$$

with ϵ_i^1 and ϵ_i^0 being identically and independently distributed additive errors. This in turn yields the following structural model of heterogeneous treatment effects approximating subject-specific effects by a covariate Z_i:

$$Y_i^a = \alpha_0 + (\alpha_1 - \alpha_0)a + \beta_0 Z_i + (\beta_1 - \beta_0)aZ_i + \epsilon_i^0 + \left(\epsilon_i^1 - \epsilon_i^0\right)a. \tag{5}$$

We assume that the variables Y_i^1, Y_i^0, and Z_i follow a trivariate normal distribution given by

$$\mathrm{N}\left[\begin{pmatrix} \mu_1 \\ \mu_0 \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 & \rho_1\sigma_1\sigma_Z \\ \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 & \rho_0\sigma_0\sigma_Z \\ \rho_1\sigma_1\sigma_Z & \rho_0\sigma_0\sigma_Z & \sigma_Z^2 \end{pmatrix}\right],$$

where μ_j is the mean of Y_i^j, μ_Z is the mean of Z_i, 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^j and Z_i is denoted by ρ_j. The trivariate normal distribution implies that the ATE is given by

$$\Delta = \mu_1 - \mu_0.$$

Further, it also implies that the joint distribution of Y_i^j and Z_i is bivariate normal. The conditional distribution of Y_i^j given Z_i is

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

where

$$\beta_j = \rho_j \frac{\sigma_j}{\sigma_Z} \tag{6}$$

$$\alpha_j = \mu_j - \beta_j \mu_Z \tag{7}$$

$$\sigma^2_{j|Z} = \sigma^2_j (1 - \rho^2_j) \tag{8}$$

hold. Generally, we cannot measure 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, $E[Y_i^j | Z_i]$, by a linear relationship described by the intercept α_j and the slope β_j. Based on the linear relationship, the expected difference $E[Y_i^1 - Y_i^0 | Z_i]$, can be estimated. If the difference $E[Y_i^1 - Y_i^0 | Z_i]$ 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.

The expected conditional ITE $E[Y_i^1 - Y_i^0 | Z_i]$ could of course be estimated via the standard linear model approach by including a biomarker–treatment interaction term and predicting the outcome under both treatments. However, this approach is not sufficient, as the variance of the ITEs is dependent on the correlation between the PRs. As this correlation is unknown, we propose a different approach which allows us estimating this parameter. Futhermore, having an estimate of this correlation, our approach allows obtaining individual response probabilities (and prediction intervals).

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. In the case of normally distributed PRs, Gadbury and Iyer (2000) and Gadbury, Iyer, and Allison (2001) rely on the trivariate normal distribution and derive bounds for the correlation coefficient ρ_{10}. 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). These bounds usually cover a wide range of ρ_{10} and are thus of limited practical value as already noticed by Lord (1955a).

The article examines a 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 (Gadbury & Iyer, 2000; Gadbury et al., 2001; Lord, 1955a). With a point estimate of the correlation between the PRs, their joint distribution can be reconstructed 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 be noted that knowledge about the correlation parameter ρ_{10} can be used for planning RCTs with less subjects if this correlation is positive. A more rapid and efficient drug development could be achieved. There is medical evidence that the assumption of independence is not reasonable when comparing treatments. In a series of articles, Cleophas et al. (Cleophas, 1996a, 1996b; Cleophas and de Vogel, 1998; Cleophas, 2000) showed by using results of crossover RCTs that a *positive correlation* ρ_{10} between the responses Y_i^1 and Y_i^0 is observed if the compared treatments 1 and 0 have a *similar pharmacological mode of action* whereas a *negative correlation* ρ_{10} between the responses Y_i^1 and Y_i^0 is observed if the compared treatments 1 and 0 each have *different pharmacological modes of action*.

2.2 | Reconstruction

We start by making assumptions about the PRs with regard to the mechanism of how the data are generated based on Cheng, Small, Tan, and Ten Have (2009, p. 21). First the PR of subject i is assumed to be independent of the treatment allocation other than that of subject i. This is commonly referred to as "stable unit treatment value assumption (SUTVA)." The second point would be that a subject enrolled in the RCT is an independent and identically distributed random sample from a well–defined population. The third assumption is that the allocation T_i and Z_i are independent as guaranteed by random allocation. Aside from these standard assumptions we relate the biomarker Z_i to the ITE.

Assume for the moment that we would know the joint distribution of the normally distributed marginals of the responses Y_i^1 and Y_i^0 as shown in Figure 3a. The diagonal line shows where the responses to the compared treatments are exactly equivalent. As described above, we make the reasonable assumption that the joint distribution of Y_i^1 and Y_i^0 follows a bivariate normal distribution.

Rotating Figure 3a by 45° in a clock-wise manner yields Tukey's well–known "sum–difference plot," that is this plot is obtained by graphing the differences $Y_i^1 - Y_i^0$ on the sums $Y_i^1 + Y_i^0$. Note that the y-axis shows the treatment effects for each

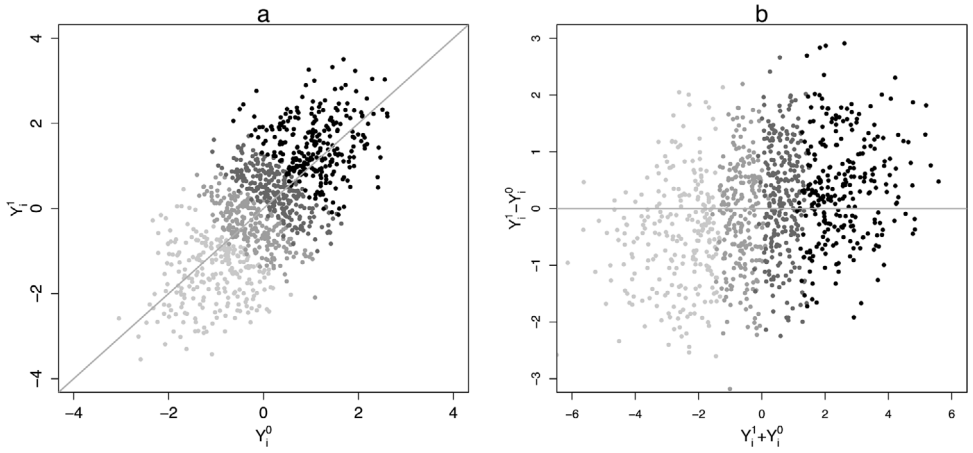

FIGURE 3 Marginal distributions Y_i^1 and Y_i^0 and "sum–difference plot." A linear relation between $Y_i^1 + Y_i^0$ and Z_i is assumed with only small measurement errors. The different gray tones denote the quantiles of Z_i. It can be seen that lighter values benefit more likely from treatment Y_i^0 while darker values benefit more likely from treatment Y_i^1

single subject. In our hypothetical example, the sum–difference plot corresponding to Figure 3a is shown in the plot of Figure 3b. The diagonal line in Figure 3a is now a horizontal line. Variation alongside the x-axis quantifies *between-subject variation* whereas variation along the y-axis quantifies *within-subject variation*. Since there is no independence between the sums $Y_i^1 + Y_i^0$ and differences $Y_i^1 - Y_i^0$, *subject–treatment interactions* are present. The between-subject variation and the subject–treatment interactions are approximated in an RCT based on the parallel-group design by a biomarker Z_i. Since the biomarker Z_i and the sum $Y_i^1 + Y_i^0$ encode for the same information, namely the between-subject variation and the subject–treatment interactions, it is reasonable to assume a functional relationship between the biomarker Z_i and the sum $Y_i^1 + Y_i^0$ given by

$$Z_i = \lambda + \kappa\left(Y_i^1 + Y_i^0\right) + \eta_i, \tag{9}$$

where λ and κ are constants and the error term η_i is assumed to be independent of Y_i^j, formally $\eta_i \perp Y_i^j$, independent of $(Y_i^1 + Y_i^0)$, formally $\eta_i \perp (Y_i^1 + Y_i^0)$ 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. If the random variables $(Y_i^1, Y_i^0, \eta_i)'$ have a normal distribution then the linear combination in (9) will be normally distributed with mean value $\lambda + \kappa(\mu_1 + \mu_0)$ and variance $\kappa^2(\sigma_1^2 + \sigma_0^2 + 2\rho_{10}\sigma_1\sigma_0) + \sigma_\eta^2$.

The use of the reconstruction variable leads to the following model of heterogeneous treatment effects:

$$Y_i^a = \alpha_0 + (\alpha_1 - \alpha_0)a + (\beta_0 + (\beta_1 - \beta_0)a)\left(\lambda + \kappa\left(Y_i^1 + Y_i^0\right) + \eta_i\right) + \epsilon_i^0 + \left(\epsilon_i^1 - \epsilon_i^0\right)a. \tag{10}$$

When looking at any model of heterogeneous treatment effects shown in (5) and (10), we can see the following restrictions.

- If $\beta_1 = \beta_0$, then no interaction between the biomarker Z_i and the treatment T_i is present. This might be due to the fact that there are indeed no heterogeneous treatment effects and thus no treatment-covariate interaction are present or due to the fact that the covariate Z_i simply cannot "recover" any heterogeneous treatment effects although indeed present. Alternatively, in case of $\beta_0 = \beta_1 \neq 0$, a prognostic biomarker Z_i is present.
- Besides, if without loss of generality $\beta_1 = -\beta_0$ holds, then no between-subject variability is present and a "pure" subject–treatment interaction is present. However, only knowledge about β_0 is necessary since the in case of $T_i = 0$ the biomarker effect is quantified by $\beta_0 Z_i$ or in case of $T_i = 1$ the biomarker effect is quantified by $-\beta_0 Z_i$.

Note that the presence of a biomarker–treatment interaction leads without loss of generality to $\sigma_1^2 - \sigma_0^2 > 0$.

The joint distribution of $(Y_i^1, Y_i^0, Z_i)'$ where Z_i is interpreted as reconstruction variable as defined by Definition (9) follows a trivariate normal distribution given by

$$\mathrm{N}\left[\begin{pmatrix} \mu_1 \\ \mu_0 \\ \lambda + \kappa(\mu_1 + \mu_0) \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 & \kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0) \\ \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 & \kappa\sigma_0(\rho_{10}\sigma_1 + \sigma_0) \\ \kappa\sigma_1(\sigma_1 + \rho_{10}\sigma_0) & \kappa\sigma_0(\rho_{10}\sigma_1 + \sigma_0) & \sigma_Z^2 \end{pmatrix}\right], \tag{11}$$

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 $\rho_{10} \in (-1, 1)$.

The proof can be found in the Supporting Information. The relation between Z_i and $Y_i^1 + Y_i^0$ can be checked in the marginal distribution plots of Z_i and Y_i^j. If a linear relation between Z_i and both of them can be assumed, then the relation between Z_i and $Y_i^1 + Y_i^0$ is also linear.

Now, the trivariate normal distribution is completely described by the reconstruction 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 reconstructing 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 are observable in parallel-group RCTs.

3 | MAXIMUM LIKELIHOOD ESTIMATION

In the following, we propose maximum likelihood estimators of the parameters for model (11) which are based on data generated in a parallel–group RCT by (Y_i, T_i, Z_i) with two treatment groups. We start with deriving maximum likelihood estimators of the parameters $(\mu_Z, \sigma_Z^2, \alpha_j, \beta_j, \sigma_{j|Z}^2)$. 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 which was first recognized by Lord (1955a) and Anderson (1957).

We start by introducing the data structure of an RCT based on the parallel-group design and corresponding notation. Let the treatment indicator T_i be independently distributed from Z_i, $T_i \perp Z_i$. Additionally, let $N = n_1 + n_0$ be the sample size with n_1 subjects allocated to treatment 1 and n_0 to treatment 0. Ordering by the treatment indicator T_i we get the following data structure (adapted from (Anderson, 1957, p. 202)) with missing data

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

with $i \in \{1, \ldots, n_1, n_1 + 1, \ldots, n_1 + n_0\}$.

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

$$\hat{\mu}_Z = \frac{1}{N}\sum_{i=1}^{N} z_i, \tag{12}$$

$$\hat{\sigma}_Z^2 = \frac{1}{N}\sum_{i=1}^{N} z_i^2 - \hat{\mu}_Z^2. \tag{13}$$

Further, the maximum likelihood estimators for β_j, α_j, and $\sigma_{j|Z}^2$ denoted by $\hat{\beta}_j$, $\hat{\alpha}_j$, and $\hat{\sigma}_{j|Z}^2$ are given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n_j}(y_i^j - \bar{y}_j)(z_i - \bar{z}_j)}{\sum_{i=1}^{n_j}(z_i - \bar{z}_j)^2}, \tag{14}$$

$$\hat{\alpha}_j = \bar{y}_j - \hat{\beta}_j \bar{z}_j, \tag{15}$$

$$\hat{\sigma}_{j|Z}^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i^j - \bar{y}_j)^2 - \hat{\beta}_j^2 \frac{1}{n_j}\sum_{i=1}^{n_j}(z_i - \bar{z}_j)^2, \tag{16}$$

where $\bar{y}_1 = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i^j$ and $\bar{z}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} z_i$.

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

$$\mu_j = \alpha_j + \beta_j \mu_Z, \tag{17}$$

$$\sigma_j^2 = \sigma_{j|Z}^2 + \beta_j^2 \sigma_Z^2, \tag{18}$$

$$\rho_j = \frac{\beta_j \sigma_Z}{\sqrt{\sigma_{j|Z}^2 + \beta_j^2 \sigma_Z^2}}. \tag{19}$$

Substituting the parameters of θ_M by the corresponding estimators for $\theta_C = (\mu_Z, \sigma_Z^2, \alpha_j, \beta_j, \sigma_{j|Z}^2)'$ provided in (12)–(16) gives the maximum likelihood estimators for θ_M denoted by

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

Similar versions of the estimates $\hat{\theta}_C$ and $\hat{\theta}_M$ were derived by Lord (1955a), Lord (1955b) and Anderson (1957).

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

The parameters for μ_j and σ_j^2 are already given in (17)–(18). The parameters $(\rho_{10}, \kappa, \lambda, \sigma_\eta^2)'$ are given by

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

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

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

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

Substituting the parameters θ_R by the estimators for θ_C provided in (12)–(16) 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 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 Supporting Information.

Besides, some restrictions appear in the parameter space and are given in terms of θ_C by

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

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

or in terms of θ_M by

$$\left(\rho_1^2 > \rho_0^2\right) \quad \wedge \quad \left(\sigma_1^2(1 - \rho_1^2) > \sigma_0^2(1 - \rho_0^2)\right) \quad \vee$$

$$\left(\rho_1^2 < \rho_0^2\right) \quad \wedge \quad \left(\sigma_1^2(1 - \rho_1^2) < \sigma_0^2(1 - \rho_0^2)\right).$$

The proof for deriving these restrictions can be found in Laubender (2014). In practice, checking the regression lines for both outcomes Y^0 and Y^1 on the biomarker Z will give an impression, whether there is an interaction effect.

In order to understand the constraints $\rho_0^2 - \rho_1^2 < 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0$, assume without loss of generality that we want to develop a cDx Z_i for a new treatment 1. A high response is more favorable and the variance σ_1^2 is higher than the variance σ_0^2 indicating the presence of subject–treatment interactions (see, e.g., 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–biomarker interaction can be established. However, as outlined in Section 1, it is not sufficient to only consider the mean of the ITEs but the variance of the ITEs (either unconditional or conditional on Z_i) should also be taken into account. The restriction of different sized residual variances $\sigma_{0|Z}^2$ and $\sigma_{1|Z}^2 < 0$ is a consequence of the structural model of heterogeneous treatment effects shown in (5) and (10).

When looking at the ITEs conditional on Z_i, three types of (qualitative) interactions can be distinguished and examples are 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 high values of Z_i, we see that the conditional mean of treatment 1 is higher than that of treatment 0, and higher responses under treatment 1 compared to 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 high values of Z_i we can now see that the conditional mean of treatment 0 is higher than under treatment 1, but under treatment 1 more higher responses are achieved 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 0 is higher than under treatment 1 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 the higher the values of response under treatment 0 are even though the mean of Y_i^1 conditional on Z_i is higher than the mean of Y_i^0 conditional on Z_i.

These constraints also imply the following well-known situations of linear regression modeling where no estimation of the reconstruction 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 without any simultaneous subject effects. This finding is in accordance with the statement that subject–treatment interactions "cannot be estimated separately from variation among the units" (Cox & Reid, 2000, p. 20). An interaction without main effect for the biomarker Z_i is present, from the point of view of the linear regression model.

4 | DATA EXAMPLE

A classic example of an RCT is taken from Schwenke (1990, table 1). The RCT randomized 24 patients equally to three types of exercise programs. The clinically relevant endpoint is the heart rate observed after treatment. 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. It shows 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 ($T_i = 0$).

The crossing of the both lines of exercises 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. A first check shows that the restrictions on the parameter space $\Theta_C^\star$ are fulfilled. 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.

Table 1 also shows the estimates of the reconstruction parameters with asymptotic 95% confidence intervals. Exercise 1 is on average lower than exercise 3 with an ATE $\hat{\Delta} = -5.75$ and the corresponding 95% confidence interval includes 0. In contrast,

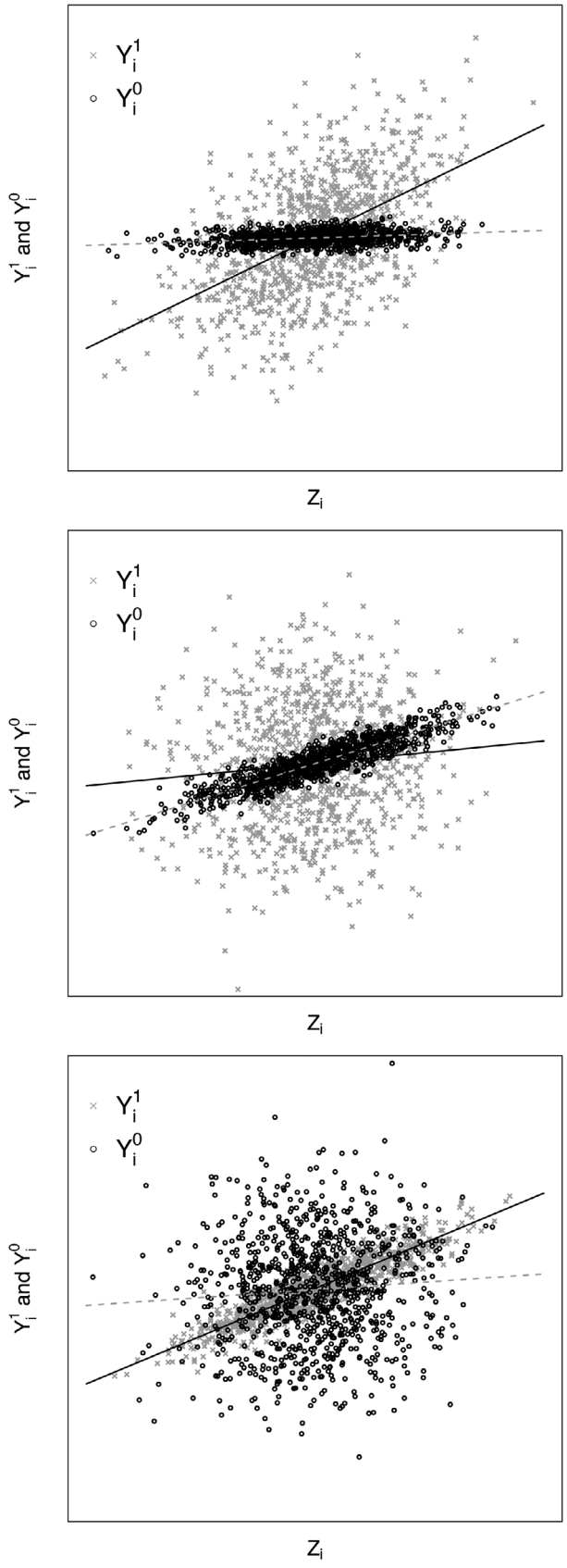

FIGURE 4 Scatter plots of simulated responses Y_i^j and a biomarker Z_i for treatment 1 (gray crosses) and for treatment 0 (black circles) with regression lines superimposed. All three plots show a qualitative interaction. Only the top plot shows data which fulfill the constraints $\rho_0^2 - \rho_1^2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

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 versus 3 ($N = 16$)		Exercise 2 versus 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 reconstruction 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
$\hat{\sigma}^2_\eta$	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
$\Phi[\hat{\Delta}/\hat{\sigma}_\Delta]$	0.35	0.14, 0.56	0.77	0.60, 0.95

exercise 2 is on average higher than exercise 3 with $\hat{\Delta} = 7.20$, here the corresponding 95% confidence interval does not include 0. However, the ATE does not inform 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. This indicates the presence of subject–treatment interactions.

Based on the estimates of the reconstruction parameters θ_R, both bivariate normal densities are drawn for both contrasts in Figure 5 (top row). From this reconstructed 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: 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 $\Phi[\hat{\Delta}/\hat{\sigma}_\Delta]$, 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, $\Phi[\hat{\Delta}/\hat{\sigma}_\Delta]$ 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 $\Phi[\hat{\Delta}/\hat{\sigma}_\Delta] = 0.77$ is observed. A similar reasoning holds for contrast B. Although the ATE is in this case in clear favor for exercise 3, 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 not only have estimates with Z_i "integrated out" like $\hat{\Delta}$ and $\Phi[\hat{\Delta}/\hat{\sigma}_\Delta]$ but also to have estimates conditional on $Z_i = z_i$. Such estimates are shown in Figure 5 (bottom row) where the treatment effects conditional on Z_i are shown with 95% confidence and prediction intervals. Here, 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 exercise 3 illustrating the qualitative nature of the treatment–biomarker interaction whereas for contrast B exercise 3 should be recommended for patients with higher values of the baseline heart rate.

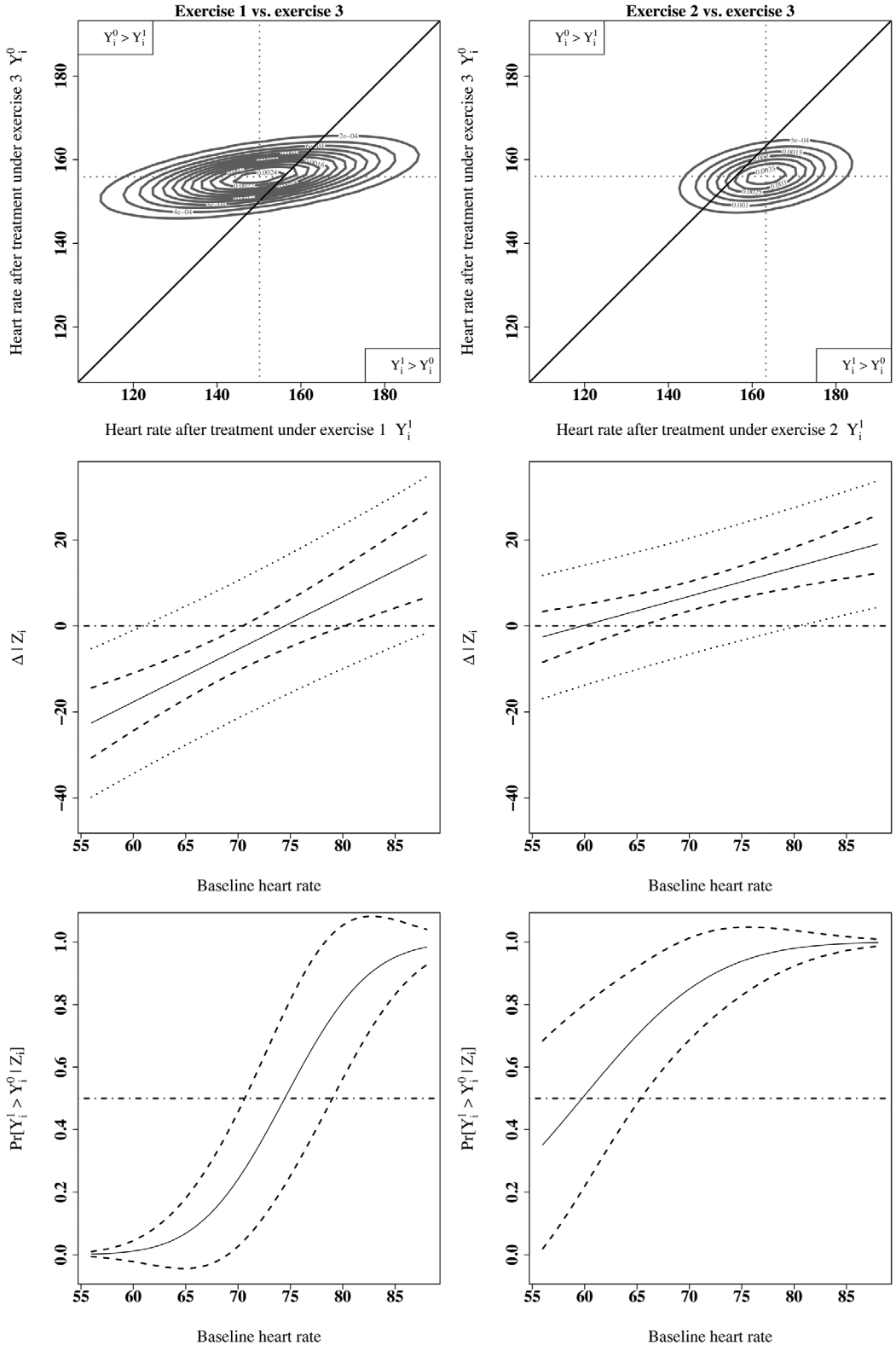

FIGURE 5 The left column of plots refers to contrast A whereas the right column of plots refers to contrast B. Top row: reconstructed joint distribution of heart rate after treatment Y_i^1 and Y_i^0 where the corresponding values of μ_1 and μ_0 are added by vertical and horizontal dotted lines. Further, 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

5 | DISCUSSION

5.1 | Conclusions

The knowledge of ITEs is helpful to select the best of two treatments for an individual subject. To improve the assessment of uncertainty of ITE estimates it is necessary to know the counterfactual correlation between the PRs Y_i^1 and Y_i^0. In a parallel–group RCT information of the marginal distributions of the single PR components is available, but no information of their joint distribution due to the "fundamental problem of causal inference" (Holland, 1986, p. 947).

We present a counterfactual approach for estimating the joint distribution of two normally distributed responses to two treatments. We assume a bivariate normal joint distribution for the PR. Additionally, we assume the presence of a normally distributed baseline biomarker Z_i which is 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 an RCT : the between-subject variation.

The biomarker Z_i as an indicator for the sum $Y_i^1 + Y_i^0$, which creates constraints for the estimation of the bivariate PR distribution. To perform successfully a maximum likelihood estimation of the reconstruction parameters θ_R, the constraints $\rho_0^2 - \rho_1^2 < 0$ and $\sigma_{0|Z}^2 - \sigma_{1|Z}^2 < 0$ have to be fulfilled. The quantity ρ_j^2 can be interpreted as proportion of the variance of Y_i^j explained by the biomarker Z_i. Similarly, the quantity $\sigma_{j|Z}^2$ can be regarded as unexplained or residual variance.

Lord (1955a) formulated this problem in the early 1950s and described the estimation of parameters in the given specific incomplete data setting. He used a trivariate normal distribution. Our model combines this idea and ideas proposed by Gadbury and Iyer (2000). This paper derives bounds of the correlation ρ_{10}. For the first time, we provide a point estimate of the correlation between the individual potential responses ρ_{10} together with a confidence interval.

Based on the more developed theory, our approach facilitates a more–informed assessment of a biomarker's relevance for treatment selection than the classical approach of estimating an interaction effect between marker and treatment. As Huang, Gilbert, and Janes (2012) have demonstrated, a strong interaction coefficient is important for a biomarker to have value for treatment selection but is not useful for summarizing its predictive performance . The predictive performance 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).

We present the dependency between biomarker value and ITE by plots of the response function $f[z_i] = \Pr[\Delta > 0 | Z_i = z_i]$. This provides a useful decision to guide individual decisions.

Huang et al. (2012) introduce an 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 $\mathrm{TPF}[z_i] = \Pr[Z_i > z_i | \Delta > 0]$, and the false positive fraction $\mathrm{FPF}[z_i] = \Pr[Z_i > z_i | \Delta < 0]$. The ROC curve is given by $\mathrm{ROC}[t] = \mathrm{TPF}[\mathrm{FPF}^{-1}[t]]$. 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 discusses the issue in the context of a randomized trial. It also stresses that the statistical interaction between marker value and treatment assigned is the primary measure of marker performance (Buyse, 2007; Freidlin & Simon, 2005; Sargent, Conley, Allegra, & Collette, 2005; Simon, 2008; Simon, Paik, & Hayes, 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 numerical interaction coefficient estimates but show different capacity in terms of classifying a subject according to treatment effectiveness. Therefore, we see the response function as important information to be communicated to clinicians (see Figure 5).

5.2 | Limitations and issues for future research

The main limitation of our approach consists in its distributional and structural assumptions: linear relationship between biomarker and PRs and the distributional assumptions, especially that the joint distribution of the PRs follows a bivariate normal PR distribution. We explore copula models as alternative to the bivariate normal PR distribution. They allow marginal normal distributions of the biomarker Z_i and the PRs Y_i^1 and Y_i^0 as well as a bivariate normal distributions of (Y_i^1, Z_i) and of (Y_i^0, Z_i) to be normal. Conditional copula models may have the potential to work out such distributions (Veraverbeke, Omelka, & Gijbels, 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.

It is well-known that univariate normality of Y_i^1, Y_i^0, and Z_i and bivariate normality of Y_i^1 and Z_i and of Y_i^0 and Z_i does not necessarily imply trivariate normality, however it is usually sufficient to evaluate these assumptions as outlined in standard text books on multivariate statistics: "in practice, …, the presence of joint nonnormality is likely to be detected quite often by methods directed at studying the marginal normality of the observations on each variable" (Gnanadesikan, 1977, p. 168) and "many types of nonnormality are often reflected in the marginal distributions and scatterplots. Moreover, for most practical work, one-dimensional and two-dimensional investigations are ordinarily sufficient. Fortunately, pathological data sets that are normal in lower-dimensional representations but nonnormal in higher dimensions are not frequently encountered in practice" (Johnson & Wichern, 1992, p. 153).

A similar critical aspect consists in the distributional assumptions of the reconstruction 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, e.g., 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 depends on a biomarker value. This aspect is not studied so far and is an issue for future research. Finally, the model is developed for univariate biomarkers. It is an issue for future research to generalize to the use of multiple biomarkers simultaneously. Attempts have been made by Kaiser and Gadbury (2013). It is also of interest to apply the basic idea of this paper to other common clinical outcomes like binary endpoints and count data, especially to remove restrictions as introduced by Huang et al. (2012) to make the counterfactual model identifiable.

The extension of our work to multivariate measurements per subject, as in longitudinal data, are of interest. Another potential extension is the analysis of time-to-event data in the described context. The concept described herein could also be applied when Y_i^1 and Y_i^0 are (potentially right-censored) survival times that follow a log-normal distribution. However, for right-censored survival times, values would have to be imputed. This is a topic for future work. Log-normally distributed event times and handling right censoring with the EM-algorithm allows to translate the event data problem in the setting of our approach. The interesting question is, how many real data settings do fit in this structural straightjacket.

5.3 | Summary

The presented work follows a line of ideas which started with Lord's paper on the calibration of scores, Anderson's likelihood approach to Lord's problem, and the work of Gadbury and colleagues to infer information on the correlation between the individual potential responses ρ_{01}. We complete this line of thinking by providing a structural model with a maximum likelihood estimation of ρ_{01} and related measures of uncertainty. Having presented the formal machinery, we demonstrate its application for individualized treatment decisions and derive decision tools for a straightforward interpretation of the formal outcomes. We stress the difference between our counterfactual and the usual approach, which combines exchangeability arguments and regression models with interaction effects between treatment and biomarker. Our approach exploits strong formal assumptions: trivariate normal distribution as well as linear relationships between biomarker and outcome. We present an example, which fit this formal setting and discuss strategies to assess these assumptions. The presented structural model offers approaches to causal sensitivity analyses of clinical trial data. There are options to generalize our model by using copula and monotone relationships between treatment outcome and biomarker. It is also of interest to explore Bayesian approaches to handle issues of non-identifiability, which easily raise for more general potential response models or when modeling non-Gaussian types of outcomes. We see our work as a proof of principle and as encouragement to explore the relevance of more general potential response models in clinical research.

ACKNOWLEDGEMENT

The text of this article is based on the dissertation of R. P. Laubender (Laubender, 2014).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Ulrich Mansmann https://orcid.org/0000-0002-9955-8906
Michael Lauseker https://orcid.org/0000-0002-6662-7127

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.

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

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

Cheng, J., Small, D. S., Tan, Z., & 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). Criticism of cardiovascular studies with negative results due to a negative correlation. *Angiology*, *47*, 139–147.

Cleophas, T. J. M. (1996b). 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., & 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, CA: Wadsworth Advanced Books and Software.

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

Freidlin, B., & 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., & Iyer, H. K. (2000). Unit–treatment interaction and its practical consequences. *Biometrics*, *56*, 882–885.

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

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Gneiting, T., & 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, NJ: Prentice Hall.

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

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

Hudis, C. A. (2007). Trastuzumab: Mechanism of action and use in clinical practice. *New England Journal of Medicine*, *357*, 39–51.

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

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

Kaiser, K. A., & Gadbury, G. L. (2013). Estimating the range of obesity treatment response variability in humans: Methods and illustrations. *Human Heredity*, *75*, 127–135.

Laubender, R. P. (2013). *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*. Munich, Germany: Verlag Dr. Hut.

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

Lord, F. M. (1955b). Equating test scores: A maximum likelihood solution. *Psychometrika*, *20*, 193–200.

Sargent, D. J., Conley, B. A., Allegra, C., & 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.

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

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

Simon, R. M., Paik, S., & 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., & Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, *38*, 766–780.

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

How to cite this article: Laubender RP, Mansmann U, Lauseker M. Estimating the distribution of heterogeneous treatment effects from treatment responses and from a predictive biomarker in a parallel-group RCT: A structural model approach. *Biometrical Journal*. 2020;62:697–711. https://doi.org/10.1002/bimj.201800370