Evaluation of Biomarkers for Pharmacological Activity

Peer-reviewed author version

ALONSO ABAD, Ariel; MOLENBERGHS, Geert; RENARD, Didier & Darstein, Christelle (2009) Evaluation of Biomarkers for Pharmacological Activity. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 19(2). p. 256-272.

DOI: 10.1080/10543400802622386
Handle: http://hdl.handle.net/1942/9290

Reference (Published version):
Alonso Abad, Ariel; Molenberghs, Geert; Renard, Didier & Darstein, Christelle(2009) Evaluation of Biomarkers for Pharmacological Activity. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 19(2). p. 256-272

DOI: 10.1080/10543400802622386
Handle: http://hdl.handle.net/1942/9290

Evaluation of Biomarkers for Pharmacological Activity

Ariel Alonso, Geert Molenberghs

Center for Statistics, Hasselt University,

Agoralaan, B–3590 Diepenbeek, Belgium.

Didier Renard

Novartis Pharma AG, Lichtstrasse 35, Basel, Switzerland.

Christelle Darstein

Lilly Services S.A., Lilly Corporate Center, Indianapolis, USA

Abstract

In recent years, the cost of drug development has increased the demands on efficiency in the selection of suitable drug candidates. Biomarkers for efficacy and safety could be a plausible strategy to improve this selection process. In the present work, we focus on the study and evaluation of different physiological variables as biomarkers for pharmacological activity. We proposed three different approaches using multivariate and univariate techniques. We note that even though one could argue that the multivariate procedure is more powerful than the other alternatives, the univariate methods also offer a great flexibility to answer interesting scientific questions. The three approaches were used to analyze a crossover study involving an opioid antagonist.

Keywords: biomarkers, surrogate marker, crossover, optimization.

1 Introduction

The rising costs of drug development and the challenge of facing new and re-emerging diseases are putting considerable demands on efficiency in the selection of suitable drug candidates. An effective strategy in improving this process is the proper selection and application of biomarkers for efficacy and safety during the different stages of the drug development pipeline.

Some authors refer to a biological marker or biomarker as a variety of physiological, pathological, or anatomical measurements that are thought to relate to some aspect of a healthy or pathological process (Temple 1995, Lesko and Atkinson 2001). In the same vein, a biomarker has also been defined as a characteristic that can be measured and evaluated as an indicator of healthy biological processes, pathological processes or pharmacological responses to therapeutic intervention (NIH Biomarker Definitions Working Group, 2001). Other definitions have since emerged and the discussion on what biomarkers should be and where to apply them continues. Biomarkers are currently being used in various areas, including disease identification, target discovery and validation, volunteer/patient inclusion and stratification during clinical studies, drug efficacy and safety and prediction of drug response (Suico *et al* (2006). Such biomarkers include measurements that help identifying the etiology of certain medical problem or the progress of a disease. They also include measurements related to the mechanism of response to treatments and actual clinical responses to therapeutic interventions (Burzykowski, Molenberghs, and Buyse 2005).

From a regulatory perspective, a biomarker is not considered an acceptable endpoint for the determination of efficacy of new drugs unless it has been shown to function as a valid indicator of clinical benefit, i.e., unless it is a valid surrogate. The NIH Biomarker Definitions Working Group (2001) also addressed the relation-

ship between biomarkers, clinical endpoints, and surrogate endpoints. A clinical endpoint is considered the most credible indicator of drug response and is defined as "a characteristic or variable that reflects how a patient feels, functions, or survives". During clinical trials, clinical endpoints should in principle be used, unless a biomarker is available that has risen to the status of a surrogate endpoint and is expected to predict either, clinical benefit, harm, or lack of both benefit and harm. Realistically, the working group points out that probably only a few biomarkers are likely to achieve a consensus surrogate endpoint status.

Biomarkers differ in their closeness to the intended therapeutic response or clinical benefit. Some biomarkers can be thought to be valid surrogates for clinical benefits, such as, for example, blood pressure or cholesterol, while they can also reflect the pathological process and could be considered potential surrogate endpoints, such as, for example, brain appearance in Alzheimer brain infarct size. The evaluation of a biomarker as potential surrogate endpoint has received considerable attention over the last decade and a detailed discussion of the main contributions in this area can be found in Burzykowski, Molenberghs, and Buyse (2005).

Additionally, other biomarkers have a more uncertain relation to clinical outcome but they can still reflect the drug action, such as, for example, angiotensin converting enzyme (ACE) inhibition, degree of binding to a receptor, or inhibition of an agonist. In the present work, we focus on the evaluation of this type of biomarkers. More specifically, emphasis will be on the evaluation of biomarkers of pharmacological activity for a certain compound.

In Section 2, we introduce the motivating case study. Section 3 covers important aspects of the analysis of crossover trials with repeated measurements. Three methods to evaluate biomarkers for pharmacological activity are introduced in Section 4. In

Section 5, a targeted simulation study is carried out to compare the relative performance of the methods introduced in Section 4. Finally in Section 6 the case study is analyzed.

2 Case Study

The case study is a three-period (P1, P2, and P3), two-treatment cross-over trial in which 15 male subjects received either Naltrexone (A) or a matching placebo (B) in a given period. An ABB–BAA design was used and, within every period, the treatment was administered at three consecutive day: D_{11}, D_{12}, and D_{13} in the first period P1; D_{21}, D_{22}, and D_{23} in the second period P2; and D_{31}, D_{32}, and D_{33} in the third period P3. Details can be found in Suico *et al* (2006).

Naltrexone is an opioid receptor antagonist, i.e., it acts by blocking the activation of opioid receptors. The goal of the study was to identify the best biomarker of pharmacological activity for this kind of compound. Several biomarkers were considered in the study: essentially, a group of variables was measured under different conditions and at two different days. At day 1, measurements of 5 neurohormones: Adrenocorticotropic (Acth), Cortisol, Luteinizing (LH), Follicle-stimulating hormone (FSH), and Prolactin were taken following a single oral dose administration of either Naltrexone or placebo. Moreover, measurements of pupil diameter under three different light conditions: scotopic (low luminosity — 0.04 lux), mesopic Lo (medium luminosity — 0.4 lux), and mesopic Hi (high luminosity — 4.0 lux) were also taken. Both pupillary diameter and neurohormones were assessed at approximately 0.5, 1, 1.5, 2, and 3 hours post-dose. At day 3, a two-minute fentanyl dose infusion was administered to all subjects 1h after receiving their oral dose of Naltrexone or placebo. Measurements of neurohormones and pupil diameter were evaluated at 1 h 20, 1 h 40, 2 h, 2 h 20, 2 h 40, and 3 h, following the oral dose of Naltrexone or placebo.

Finally, at approximately 3 h 10 post oral dose, a cold pressor test was administered after completion of the pupillometry and neurohormones measurements. The cold pressor test (Cp) is typically used to evaluate the analgesic effects of a compound (such as fentanyl). Thereupon, this test will show the ability of Naltrexone to block the effects of fentanyl. The test consisted in rating the pain felt by a subject during 2 minutes following immersion of the subject's dominant hand in warm and cold baths. During the test, participants verbally rated pain intensity at time 0, and at 15, 30, 60, 90, and 120 seconds.

Day 1 measurements were taken following a single dose of Naltrexone or placebo and therefore, they represent the direct pharmacological action. Day 3 measurements followed a short infusion of the opioid receptor agonist fentanyl, that is, a substance that increases the activation of opioid receptors. Hence, measurements taken at the third day show the ability of Naltrexone, an antagonist, to block the pharmacological effects of the agonist. We can then consider that these measurements represent another way of evaluating the pharmacological activity of the antagonist compound.

The previous discussion suggests the analysis of the three day 1 measurements, taken at D_{11}, D_{21}, and D_{31}, as an initial cross-over study and then, separately the three day 3 measurements, taken at D_{13}, D_{23}, and D_{33}, as a further cross-over study.

The variables recorded for each subject are displayed in Table 1. Eight biomarkers were measured at the first and third days whereas one (Cp) was measured only at day three. Each combination biomarker-day is of scientific interest and hence 17 responses in total will be analyzed.

The main objective of the study is to identify biomarkers for pharmacological activity and therefore, we are primarily interested in determining for which biomarker the difference between Naltrexone and placebo is largest.

Table 1: *Candidate biomarkers.*

Biomarkers	Day 1	Day 3
Acth	$\star$	$\star$
Cortisol	$\star$	$\star$
LH	$\star$	$\star$
FSH	$\star$	$\star$
Prolactine	$\star$	$\star$
Mesopic Hi	$\star$	$\star$
Mesopic Lo	$\star$	$\star$
Scotopic	$\star$	$\star$
Cp		$\star$

3 Cross-over Designs and Repeated Measurements: Remarks for the Analysis

This was a cross-over study with repeated measurements in which two treatments, three periods and two sequences were considered: ABB and BAA. This design is optimal in the sense that it allows for a minimum variance unbiased estimator for the treatment effect.

Let us denote by $Y_{ih\ell j}$ the response observed at time point j for the ℓ^{th} subject in period h and in sequence group i. Additionally, we shall denote by t_O, p, s, and m the number of treatments, periods, sequences and time points, respectively. Note that for the ABB–BAA design, $t_O = 2$, $p = 3$, and $s = 2$. Further, we shall define

$$\overline{Y}_{ih.j} = \frac{1}{n_i} \sum_{\ell=1}^{n_i} Y_{ih\ell j},$$

where n_i is the number of patients in sequence group i. According to Jones and Kenward (2003) one of the issues in modeling cross-over data with repeated measurements is how best to handle both the between-period and within-period covariance structure. These authors observed that, in the two sequence design, one can

avoid the need to introduce a between-period structure by exploiting the fact that all estimators take the form $\overline{A}_{1j} - \overline{A}_{2j}$, where $\overline{A}_{ij} = \sum_{h=1}^{p} a_h \overline{Y}_{ih.j}$ and, for within-subject estimators, $\sum_{h=1}^{p} a_h = 0$. Conventional repeated measurement methods can then be applied to the derived subject contrast

$$C_{i\ell j} = \sum_{h=1}^{p} a_h Y_{ih\ell j}.$$

For the design considered in this study, i.e., ABB–BAA, one can estimate the treatment effect using the contrast

$$CT_{i\ell j} = -2Y_{i1\ell j} + Y_{i2\ell j} + Y_{i3\ell j}. \tag{1}$$

Indeed, if the previous contrast is calculated for each subject (within sequence), then the treatment effect can be expressed as the difference between the contrast's mean values in each of the two sequence groups. This allows us to evaluate the treatment effect by applying classical repeated-measurement modeling techniques to the previously defined contrasts, without having to introduce a between-period structure in our models.

In the subsequent analyses, we shall consider the following longitudinal model

$$CT_{i\ell j} = \mu_{ij} + \varepsilon_{i\ell j}, \tag{2}$$

where μ_{ij} denotes the mean for the i^{th} sequence at the j^{th} time point, with $i = 1, 2$ and $j = 1, \ldots, m$, and $\boldsymbol{\varepsilon}_{i\ell} = (\varepsilon_{i\ell 1}, \varepsilon_{i\ell 2} \ldots, \varepsilon_{i\ell m})'$ denotes a vector of error terms which is assumed to follow an m-dimensional normal distribution with mean zero and unstructured variance-covariance matrix Σ. Note that the previous model describes the mean structure using a parameter for each sequence group by time combination and, therefore, the mean structure is modeled in a saturated way. Likewise, the covariance structure is modeled using a fully general unstructured matrix.

Using the previous modeling framework, in the next section, we shall introduce three possible methods to determine the best biomarker of pharmacological activity.

4 Three Strategies for the Selection of the Best Biomarker

4.1 Approach I: The Ellipsoid Method

Following the notation introduced in Section 3, let us denote the mean evolution over time for the i^{th} sequence by $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})'$. We shall further denote by $\widehat{\boldsymbol{\mu}}_i$ the maximum likelihood estimator for the previous mean profile, computed based on the saturated linear model (2). The treatment effect over time $\boldsymbol{\Delta}_T = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ can then be estimated as $\widehat{\boldsymbol{\Delta}}_T = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2$, where $\widehat{\boldsymbol{\Delta}}_T$ has asymptotic distribution $\widehat{\boldsymbol{\Delta}}_T \sim N(\boldsymbol{\Delta}_T, \Sigma_{\Delta T})$. Note that $\Sigma_{\Delta T}$ can be consistently estimated using the hessian matrix obtained from model (2). From the previous results and using the Cochran theorem we have that, asymptotically

$$(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T)'\widehat{\Sigma}_{\Delta T}^{-1}(\widehat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}_T) \sim \chi^2_m,$$

where $\widehat{\Sigma}_{\Delta T}$ is the consistent estimator of $\Sigma_{\Delta T}$. This leads to the confidence region

$$R = \{\boldsymbol{\Delta}_T : (\boldsymbol{\Delta}_T - \widehat{\boldsymbol{\Delta}}_T)'\widehat{\Sigma}_{\Delta T}^{-1}(\boldsymbol{\Delta}_T - \widehat{\boldsymbol{\Delta}}_T) \leq C(\alpha)\},$$

where the constant is chosen so that $P(R) = 1 - \alpha$. Testing the hypothesis $H_0 : \boldsymbol{\Delta}_T = 0$ can now be done by verifying whether $0 \in R$ or, equivalently, using the test $\widehat{\boldsymbol{\Delta}}_T'\widehat{\Sigma}_{\Delta T}^{-1}\widehat{\boldsymbol{\Delta}}_T > C(\alpha)$. Further, let us denote by r the distance between zero and the ellipsoid defined by the frontier of R

$$\partial R = \{\boldsymbol{\Delta}_T : (\boldsymbol{\Delta}_T - \widehat{\boldsymbol{\Delta}}_T)'\widehat{\Sigma}_{\Delta T}^{-1}(\boldsymbol{\Delta}_T - \widehat{\boldsymbol{\Delta}}_T) = C(\alpha)\}.$$

Note that r is the solution of the optimization problem

$$r = \min_{\boldsymbol{\Delta}_T \in \partial R} \| \boldsymbol{\Delta}_T \|^2 . \tag{3}$$

Similar to the univariate setting, we argue that the larger r is, the further the ellipsoid is from the origin and therefore the larger the treatment effect is. The problem is then reduced to finding the solution of the optimization problem given in (3). The following theorem offers an analytic expression for this solution.

Theorem 1 *The solution of the optimization problem (3) is given by*

$$r = \sum_i \left(\frac{q_i \lambda}{\alpha_i + \lambda} \right)^2, \tag{4}$$

where

a) α_i *are the eigenvalues of* $\widehat{\Sigma}_{\Delta T}$,

b) $\boldsymbol{q}' = (q_1, q_2, \ldots, q_m) = P\widehat{\boldsymbol{\Delta}}_T$ *with* P *an orthogonal matrix so that* $\widehat{\Sigma}_{\Delta T} = P' D_0 P$, *and* $D_0 = (\alpha_i)_{ii}$,

c) λ *is a root of*

$$\sum_i \frac{\alpha_i q_i^2}{(\alpha_i + \lambda)^2} = C(\alpha).$$

An outline of the proof can be found in the appendix. Under Approach I, one can calculate, for each biomarker, the distance from zero to the corresponding ellipsoid and then choose as the best biomarker the one for which its ellipsoid is furthest away from the origin. Note that this last step is of a univariate type, even though the input for it is derived from multivariate calculations.

4.2 Approach II: The L_2-Norm Method

Let us start by considering the following model

$$\begin{cases} Y_1(t) = f_1(t) + \varepsilon_1(t), \\ Y_2(t) = f_2(t) + \varepsilon_2(t), \end{cases}$$

where t denotes time and Y_i the response variable for group i ($i = 1, 2$), f_i is a general function that describes the average time evolution for group i, and $(\varepsilon_1(t), \varepsilon_2(t))$ follows a bivariate Gaussian distribution with mean zero and variance-covariance matrix $\Sigma(t)$.

In the absence of treatment effect, $f_1(t) = f_2(t)$ and therefore it is intuitively appealing to use the distance between f_1 and f_2 as a measure of the effect's magnitude. If we further denote the time interval by $I = [a, b]$, then we can measure the distance between f_1 and f_2 using the L_2-norm

$$d_2(f_1, f_2)^2 = \| f_1(t) - f_2(t) \|^2 = \int_a^b [f_1(t) - f_2(t)]^2 \, dt. \tag{5}$$

In practice, $g(t) = f_1(t) - f_2(t)$ is unknown and hence needs to be estimated. We can estimate g, for instance, through fitting a saturated linear model like (2) in such standard software packages as SAS, R, or Splus. Given that we can only consider a fixed set of time points $\{t_1, t_2, \ldots, t_m\}$, fitting a saturated model merely leads to estimates of g at these prespecified values

$$\mu_j^* = g(t_j) = f_1(t_j) - f_2(t_j) = \mu_{1j} - \mu_{2j},$$

($j = 1, \ldots, m$). Using the points (t_j, μ_j^*), we can approximate (5) through the trapezoidal integration formula

$$\| f_1(t) - f_2(t) \|^2 = \| g(t) \|^2 \approx \upsilon(f_1, f_2) = \sum_{j=1}^{m-1} \frac{\mu_j^{*2} + \mu_{j+1}^{*2}}{2} \Delta_j,$$

where $\Delta_j = t_{j+1} - t_j$. Note that $\upsilon(f_1, f_2)$ can also be written as

$$\upsilon(f_1, f_2) = \sum_{j=1}^{m} \alpha_j \mu_j^{*2} = \sum_{j=1}^{m} \alpha_j (\mu_{1j} - \mu_{2j})^2,$$

where $\alpha_j = (\Delta_{j-1} + \Delta_j)/2$ and $\Delta_0 = \Delta_m = 0$. In terms of the original time points, the weights α_j take the form $\alpha_j = (t_{j+1} - t_{j-1})/2$, with $t_0 = t_1$ and $t_{m+1} = t_m$.

If we further denote by $\widehat{\boldsymbol{\mu}}_i' = (\widehat{\mu}_{i1}, \widehat{\mu}_{i2}, \ldots, \widehat{\mu}_{im})$ the maximum likelihood estimator of $\boldsymbol{\mu}_i$, $(i = 1, 2)$, then we have that

$$\widehat{v}(f_1, f_2) = \sum_{j=1}^{m} \alpha_j (\widehat{\mu}_{1j} - \widehat{\mu}_{2j})^2. \tag{6}$$

Taking into account that $\widehat{\boldsymbol{\Delta}}_T \sim N(\boldsymbol{\Delta}_T, \Sigma_{\Delta T})$ with $\widehat{\boldsymbol{\Delta}}_T = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2$ we can apply now the delta method to obtain

$$\widehat{v}(f_1, f_2) = \sum_{j=1}^{m} \alpha_j \widehat{\Delta}_{Tj}^2 \sim N(v(f_1, f_2), \sigma^2_{N(f_1,f_2)}), \tag{7}$$

where $\sigma^2_{N(f_1,f_2)} = \boldsymbol{\delta}' \Sigma_{\Delta T} \boldsymbol{\delta}$ and $\boldsymbol{\delta} = (2\alpha_1 \Delta_{T1}, 2\alpha_2 \Delta_{T2}, \ldots, 2\alpha_m \Delta_{Tm})'$. A consistent estimator for the variance of $\widehat{v}(f_1, f_2)$ can be obtained as $\widehat{\sigma}^2_{N(f_1,f_2)} = \widehat{\boldsymbol{\delta}}' \widehat{\Sigma}_{\Delta T} \widehat{\boldsymbol{\delta}}$, where $\widehat{\boldsymbol{\delta}}$ is computed by substituting Δ_{Ti} by its maximum likelihood estimator in $\boldsymbol{\delta}$. Finally, from (7) the following confidence interval follows

$$CI_\alpha[v(f_1, f_2)] = \left[\widehat{v}(f_1, f_2) - z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{N(f_1,f_2)}, \widehat{v}(f_1, f_2) + z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{N(f_1,f_2)}\right]. \tag{8}$$

We would like to point out that $v(f_1, f_2)$ has been considered under the current Approach II as an approximation to the distance between f_1 and f_2. We are then constructing confidence intervals, not for the parameter of interest $\| f_1(t) - f_2(t) \|^2$, but rather for an approximation of this distance. If (8) contains zero, then the data are not in contradiction with the equal treatment effects hypothesis.

4.3 Approach III: Different Weights Method

In Approach II, $v(f_1, f_2)$ was considered an approximation for the L_2 distance between f_1 and f_2. However, we could consider this parameter in the following, more general, way

$$v(f_1, f_2) = \sum_{j=1}^{m} \alpha_j (\mu_{1j} - \mu_{2j})^2, \tag{9}$$

with $\alpha_j > 0$ and $\sum_j \alpha_j = 1$. By using different sets of weights one can study a variety of interesting questions, such as, for instance, for which biomarker the treatment

effect is largest at the end of the study. Alternatively, we may be interested in finding the biomarker for which the treatment effect is mainly expressed at the beginning of the study, and so on. All of these situations can be explored using (9), by selecting an appropriate set of weights. Essentially, choosing different sets of weights can allow us to *zoom in* in certain specific regions of the longitudinal profiles, presumably increasing our chances of finding an effect in that regions. Here again, we can construct confidence intervals in a similar way as we did in the previous subsection and finally we could select the biomarker with interval farthest away from the origin.

In the following section we shall further explore the previous proposals using a limited simulation study. The idea is to get a better insight about the performance of these procedures under controlled conditions, where the order between the biomarkers regarding treatment effect is known.

5 Simulation Study

In all simulations, linear mean profiles were considered for each group and three main scenarios were taken into account: 1) the mean profiles were parallel; 2) the difference between the mean profiles was largest at the beginning of the time sequence; and 3) the difference between the mean profiles was largest at the end of the time sequence. As the name indicates, in the parallel mean profiles setting $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + k \cdot \mathbf{1}$ where $\mathbf{1}^{'} = (1, 1, \ldots, 1)$ is an m-dimensional vector. The constant k was biomarker-specific, and it was chosen such that the first biomarker had the smallest corresponding treatment effect, whereas the last one had the largest associated treatment effect. In the second setting, the difference between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ was largest at the beginning of the sequence and decreased towards the end. Like before, this difference was biomarker-specific and chosen as an increasing function of the biomarker's number. Finally, in the

third scenario, the difference between the mean profiles was largest towards the end of the sequence and, here again, the difference was an increasing function of the biomarker's number. Following the case study, in all the previous settings mean profiles for 17 biomarkers were generated. Along the ideas of Section 4.2 we can quantify the treatment effect using the area (A) encircled by the mean profiles for the treated and control group. In the simulation study, these areas were given by:

- *Parallel setting*: $A_i = [1 + \log(i - 0.5)]\, A_1$, where $2 \leq i \leq 17$ denotes the biomarker and $A_1 = 10$.
- *More effect at the beginning setting*: $A_{(18-i)} = A_{17} - 12.5\,[-1 + \log(i + 1)]$, where $2 \leq i \leq 17$ denotes the biomarker and $A_1 = 10$.
- *More effect at the end setting*: $A_i = A_1 + 12.5\,[-1 + \log(i + 1)]$, where $2 \leq i \leq 17$ denotes the biomarker and $A_1 = 10$.

Once the mean profiles were specified, the error terms were generated using a multivariate normal distribution with an AR1 variance-covariance matrix, characterized by a correlation $\rho = 0.5$ and a variance parameter $\sigma^2 = 1$. In all scenarios, 6 equally space time points were used: 1, 2, 3, 4, 5, 6, and a sample size of 15 subjects was considered. For each setting, 500 data sets were created and model (2) was fitted to the generated data for each biomarker separately. Using this model, maximum likelihood estimates of the mean parameters and the variance-covariance matrix were obtained. Further, the three approaches previously introduced were applied and the biomarkers ranked.

For approach III, three different sets of weights were chosen: (a) equal weights at all time points, denoted by 'Eq' in Table 2; (b) 67% of the weight equally assigned to the first half of the longitudinal sequence, 33% equally assigned to the second

half, and denoted by 'Be' in Table 2; (c) 33% of the weight equally assigned to the first half of the longitudinal sequence, 67% equally assigned to the second half, and denoted by 'End' in Table 2.

The results of the simulations are summarized in Table 2. The table is divided into three different horizontal sections, corresponding to the three settings studied. There is a column for each of the procedures used in the analysis and each of these columns is divided into two sub-columns denoted as Rank and Power. The Rank sub-column contains the median of the ranks assigned to each biomarker by the corresponding method, whereas the Power sub-column displays the proportion of times the method detected a significant difference between both treatment groups. Essentially, it was analyzed in how many cases the confidence region used by the method did not contained zero.

Generally, all approaches seem to perform very reasonably; the median of the ranks clearly reflects the way the data were generated with the lowest values appearing for the biomarkers identified with the largest numbers. Surprisingly, no major differences were found between the methods regarding the ranking. For instance, similar results were obtained, irrespective of the set of weights used for approach III. Many reasons could be put forward to explain this behavior: perhaps more extreme sets of weights need to be used, or it is also possible that the variation over time considered in the simulations may have been relatively too small with respect to the overall treatment effect. However, regarding power, some clear differences between the methods were observed. For example, in all settings the ellipsoid method exhibited the highest power.

Table 2: *Simulation study. Median of the ranks for each biomarker.*

Biom	Ellipsoid		L_2		Eq		Be		End	
	Rank	Power	Rank	Power	Rank	Power	Rank	Power	Rank	Power
Parallel profiles.										
1	17.0	0.98	17.0	0.46	17.0	0.51	17.0	0.47	17.0	0.48
2	16.0	1.00	16.0	0.99	16.0	0.99	16.0	0.99	16.0	0.99
3	15.0	1.00	15.0	1.00	15.0	1.00	15.0	0.99	15.0	0.99
4	14.0	1.00	14.0	1.00	14.0	1.00	14.0	0.99	14.0	0.99
5	13.0	1.00	13.0	1.00	13.0	1.00	13.0	1.00	13.0	1.00
6	12.0	1.00	12.0	1.00	12.0	1.00	12.0	1.00	12.0	1.00
7	11.0	1.00	11.0	1.00	11.0	1.00	11.0	1.00	11.0	1.00
8	10.0	1.00	10.0	1.00	10.0	1.00	10.0	1.00	10.0	1.00
9	9.0	1.00	9.0	1.00	9.0	1.00	9.0	1.00	9.0	1.00
10	8.0	1.00	8.0	1.00	8.0	1.00	8.0	1.00	8.0	1.00
11	7.0	1.00	7.0	1.00	7.0	1.00	7.0	1.00	7.0	1.00
12	6.0	1.00	6.0	1.00	6.0	1.00	6.0	1.00	6.0	1.00
13	5.0	1.00	5.0	1.00	5.0	1.00	5.0	1.00	5.0	1.00
14	4.0	1.00	4.0	1.00	4.0	1.00	4.0	1.00	4.0	1.00
15	3.0	1.00	3.0	1.00	3.0	1.00	3.0	1.00	3.0	1.00
16	3.0	1.00	3.0	1.00	3.0	1.00	3.0	1.00	3.0	1.00
17	2.0	1.00	2.0	1.00	2.0	1.00	2.0	1.00	2.0	1.00
Mean difference larger at the beginning.										
1	14.0	1.00	14.0	1.00	14.0	1.00	14.0	1.00	14.0	1.00
2	14.0	1.00	14.0	1.00	14.0	1.00	14.0	1.00	14.0	1.00
3	13.0	1.00	14.0	1.00	13.0	1.00	13.0	1.00	14.0	1.00
4	13.0	1.00	13.0	1.00	13.0	1.00	13.0	1.00	13.0	1.00
5	12.0	1.00	13.0	1.00	12.5	1.00	12.0	0.99	13.0	0.99
6	11.0	1.00	12.0	1.00	11.5	1.00	11.0	1.00	12.0	1.00
7	11.0	1.00	12.0	1.00	11.0	1.00	11.0	1.00	12.0	1.00
8	11.0	1.00	11.0	1.00	11.0	1.00	11.0	1.00	11.0	1.00
9	10.0	1.00	10.0	1.00	10.0	1.00	10.0	1.00	10.0	1.00
10	9.0	1.00	9.0	1.00	9.0	1.00	9.0	1.00	9.0	1.00
11	8.0	1.00	8.0	1.00	8.0	1.00	8.0	1.00	8.0	1.00
12	7.0	1.00	7.0	1.00	7.0	1.00	7.0	1.00	7.0	1.00
13	6.0	1.00	6.0	1.00	6.0	1.00	6.0	1.00	5.0	1.00
14	5.0	1.00	4.0	1.00	5.0	1.00	5.0	1.00	4.0	1.00
15	3.0	1.00	3.0	1.00	3.0	1.00	4.0	1.00	3.0	1.00
16	2.0	1.00	2.0	1.00	2.0	1.00	2.0	1.00	2.0	1.00
17	2.0	1.00	2.0	1.00	2.0	1.00	2.0	1.00	1.0	1.00
Mean difference larger at the end.										
1	15.0	0.98	14.5	0.48	15.0	0.54	14.0	0.48	15.0	0.52
2	15.0	0.99	15.0	0.49	15.0	0.54	15.0	0.47	15.0	0.53
3	14.0	0.99	13.0	0.63	13.0	0.70	13.0	0.62	13.0	0.66
4	13.0	0.99	12.0	0.70	12.0	0.75	12.0	0.70	12.5	0.75
5	12.0	0.99	12.0	0.73	12.0	0.80	12.0	0.71	12.0	0.81
6	11.0	1.00	11.0	0.77	11.0	0.84	11.0	0.78	11.0	0.83
7	10.0	1.00	10.0	0.83	10.0	0.87	9.0	0.82	10.0	0.87
8	9.0	1.00	8.5	0.89	8.5	0.92	9.0	0.89	9.0	0.93
9	8.0	1.00	8.0	0.89	8.0	0.92	8.0	0.88	8.0	0.92
10	8.0	1.00	8.0	0.89	8.0	0.93	8.0	0.91	8.0	0.93
11	7.0	1.00	6.0	0.91	7.0	0.93	7.0	0.90	6.0	0.93
12	6.0	1.00	6.0	0.91	6.0	0.96	6.0	0.91	6.0	0.95
13	6.0	1.00	6.0	0.92	6.0	0.95	6.0	0.92	6.0	0.95
14	6.0	1.00	6.0	0.95	6.0	0.97	6.0	0.95	6.0	0.97
15	5.0	1.00	6.0	0.93	5.0	0.97	6.0	0.94	5.0	0.97
16	5.0	1.00	5.0	0.95	5.0	0.97	5.0	0.95	5.0	0.98
17	5.0	1.00	5.0	0.97	5.0	0.98	5.0	0.95	5.0	0.98

This is not a surprising result. Indeed, both the L_2-norm and weighted distance methods construct a confidence interval for $\upsilon(f_1, f_2)$, which is a summary statistic for the mean differences $\boldsymbol{\Delta}_T$. Nevertheless, the ellipsoid approach constructs a confidence region for $\boldsymbol{\Delta}_T$ in a multivariate fashion. Arguably, the loss of information derived from using a summary statistics could imply a reduction of power that may explain this finding. Moreover, a closer look at the performance of the weighted distance method for the different sets of weights reveals a very coherent behavior. For instance, when the mean profiles are parallel, setting equal weights along the entire time sequence results in a mild gain in power. The same result is observed in the last setting when most of the effect is present towards the end of the study. Under this scenario, assigning more weights at the end of the sequence results in a higher power than the one obtained, for example, with the L_2-norm.

Obviously, more simulations will be necessary to shed additional light on this specific issue. However, the information obtained from this study reinforces our confidence in the general performance of the three methods introduced in the previous sections.

6 Analysis of Case Study

We shall now apply the three approaches defined in Section 4 to the data introduced in Section 2. A logarithmic transformations was used for the neurohormones variables.

6.1 Exploratory Analysis

Let us start by noting that conventional graphical techniques for longitudinal data would ignore the cross-over design of our study. For instance, in a mean profile by treatment graph, each patient would contribute to both treatment groups, ignoring the between-period association. Hence, it is more appropriate to base our

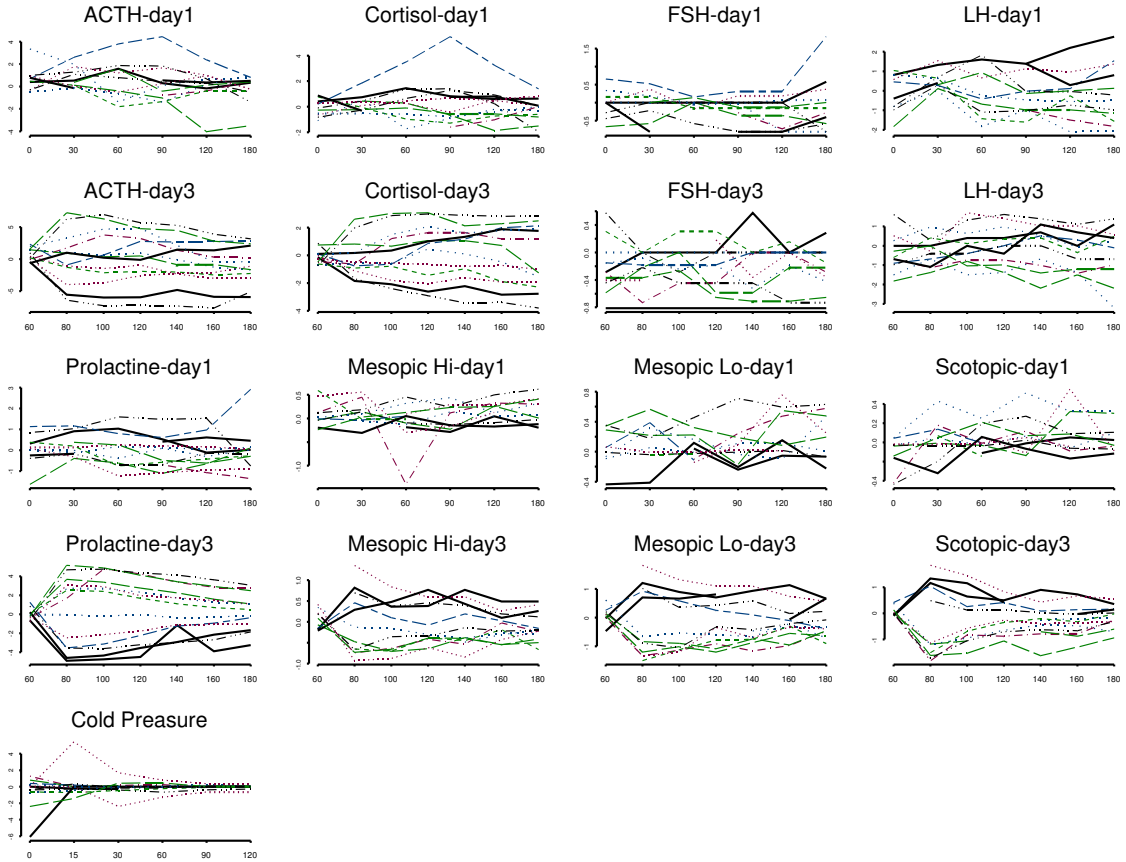

Figure 1: *Contrast profiles.*

exploratory analysis on the individual contrasts (1). The contrast profile per patient is presented in Figure 1.

The difference between the two mean contrast profiles for each sequence provides an estimate of the evolution of the treatment effect over time. This evolution is displayed in Figure 2. Note that any deviation from the horizontal zero-line indicates a treatment effect. This graph hints on a nonlinear evolution of the treatment effect over time. It also emanates from Figure 2 that Prolactine at day 3 is the biomarker in which the largest treatment effect is observed. This pattern is also present in Figure 1, where two clearly differentiated groups can be observed for Prolactine at day 3.

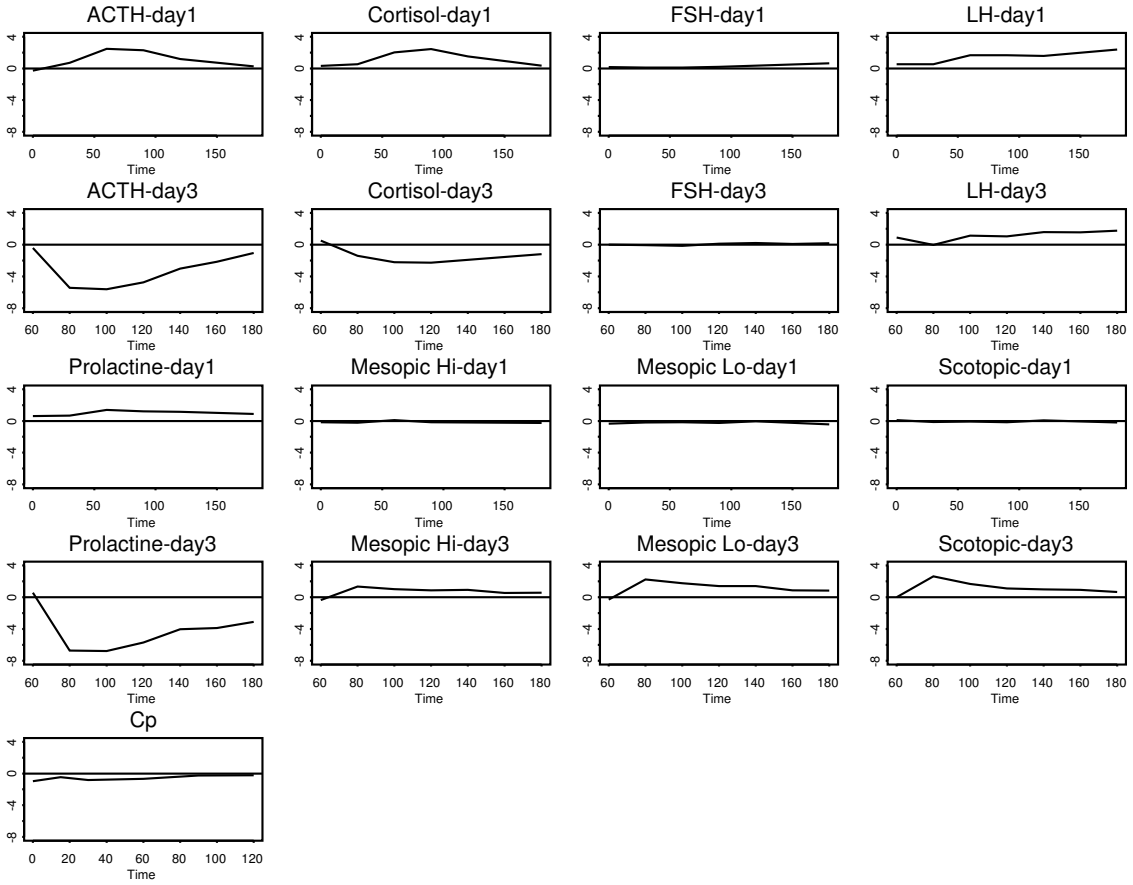

Figure 2: *Mean treatment effect over time: Naltrexone-Placebo.*

Another important issue emerging from the exploratory analysis is the difference between the relative behavior of Naltrexone and placebo at days one and three. For the biomarkers on which the treatment effect seems to be largest, the mean evolution of the Naltrexone group lies above the mean evolution of the control group at day one, provoking a positive treatment effect. However, this behavior seems to be reversed at the third day. This could be explained by the infusion of fentanyl administrated to the patients at day 3 before the measurements were taken.

Fentanyl is an opioid receptor agonist, i.e., it increases the activation of opioid receptors. This could explain the lower mean evolution of the treatment group relative to the placebo. Nevertheless, Figure 2 illustrates that, in spite of the initial

Table 3: *Ellipsoid method.*

Biomarker	r-distance
ACTH-day1	0.000
ACTH-day3	2.374
Cortisol-day1	0.003
Cortisol-day3	0.381
LH-day1	0.505
LH-day3	0.524
FSH-day1	0.000
FSH-day3	0.000
Prolactine-day1	0.129
Prolactine-day3	30.631
Mesopic Hi-day3	1.492
Mesopic Lo-day1	0.000
Scotopic-day1	0.000
Scotopic-day3	3.556
Cp	0.000

decrease in the Naltrexone effect, a tendency to recover over time appears towards the end of the time interval.

6.2 Assessment of Biomarker Quality

Given the large number of biomarkers used in the study, a Bonferroni correction was applied to account for the number of confidence intervals involved, which equals the number of biomarkers under scrutiny. Table 3 displays the results obtained after applying the procedure described in Section 4.1. Using Theorem 1, we calculated the distance from the origin to each of the ellipsoids defined by the biomarkers. For Mesopic Lo-day3 and Mesopic Hi-day1, model (2) did not converge and therefore these biomarkers were not included in the analysis.

For ACTH at day 1, FSH at days 1 and 3, Mesopic Lo at day 1, Scotopic at day 1,

and Cp, the values for the r-distance were so small that they were set to zero. Table 3 is in complete agreement with the findings of the exploratory analysis. Prolactine at day 3 was clearly the biomarker with the ellipsoid furthest away from zero, followed by Scotopic at day 3 and ACTH at day 3.

Additionally, we estimated the distance between f_1 and f_2, as described in Section 4.2. The results are summarized in the first three columns of Table 3, where LL and UL denote the lower and upper limits of the corresponding confidence interval, respectively. Here again, the Prolactine at day 3 is the clear winner, followed again by the Scototopic at day 3 and Mesopic Hi at day 3. Note that for other biomarkers, like ACTH at day 3 or Cortisol at day 3, the confidence interval contains the origin and therefore the hypotheses of no treatment effect could not be rejected in these cases. This seems to contradict the conclusions found with the ellipsoid method with which these biomarkers produced an ellipsoid that did not contain the origin.

The results obtained in the simulations can help to explain this issue. As stated in section 5 the ellipsoid approach constructs a confidence region for $\mathbf{\Delta}_T$ in a multivariate fashion whereas the other methods only work with summary statistics of this vector. The simulation study showed that the loss of information derived from using a summary statistics could imply a reduction in power to detect a treatment effect. Note also that some of these biomarkers produced very small values for the r-distance, which can also help to explain the results found when the L_2-norm was used. The ACTH at day 3, which was ranked third by the ellipsoid method, produces here a very large point estimate for $\upsilon(f_1, f_2)$ but with a very wide confidence interval that contains zero.

Finally, we analyzed the data following the approach introduced in Section 4.3. Similarly to what was done in the simulation study, in this analysis three different sets of weights were considered: (a) equal weights at all time points, denoted by 'Eq'

Table 4: *L_2 and different weights results.*

Biomarker	L_2	LL	UL	Eq	EqLL	EqUL	Be	BeLL	BeUL	End	EndLL	EndUL
ACTH-d1	22.46	-12.58	57.50	2.10	-1.12	5.31	2.05	-1.03	5.13	2.14	-1.29	5.57
ACTH-d3	128.75	-143.93	401.43	16.48	-20.47	53.43	20.30	-22.10	62.71	12.65	-18.90	44.20
Cort.-d1	24.34	-15.79	64.47	2.14	-1.30	5.58	1.92	-1.01	4.86	2.36	-1.63	6.34
Cort.-d3	22.54	-33.01	78.09	3.32	-5.51	12.14	3.59	-5.14	12.33	3.04	-5.90	11.99
LH-d1	28.62	-2.92	60.15	2.79	-0.28	5.87	2.51	-0.47	5.48	3.08	-0.25	6.40
LH-d3	9.96	-5.97	25.89	1.88	-1.16	4.92	1.61	-1.23	4.46	2.14	-1.12	5.41
FSH-d1	1.07	-2.30	4.44	0.104	-0.24	0.45	0.07	-0.18	0.33	0.13	-0.30	0.57
FSH-d3	0.10	-0.30	0.50	0.02	-0.07	0.10	0.02	-0.04	0.08	0.02	-0.09	0.13
Prol.-d1	10.59	-1.06	22.24	1.024	-0.12	2.17	0.97	-0.13	2.07	1.08	-0.15	2.30
Prol.-d3	206.03	64.68	347.37	27.56	7.91	47.20	32.17	10.18	54.17	22.93	5.57	40.30
Mes. Hi-d3	6.41	2.57	10.24	0.86	0.36	1.35	0.98	0.41	1.56	0.73	0.31	1.16
Mes. Lo-d1	0.58	-0.41	1.58	0.08	-0.05	0.21	0.07	-0.07	0.22	0.08	-0.05	0.21
Scot.-d1	0.14	-0.13	0.41	0.01	-0.01	0.04	0.01	-0.01	0.03	0.02	-0.01	0.05
Scot.-d3	16.64	6.96	26.32	2.04	0.79	3.30	2.51	1.06	3.96	1.58	0.50	2.66
Cp	38.75	-61.98	139.49	0.38	-0.79	1.56	0.46	-1.07	1.99	0.31	-0.52	1.14

LL,UL: Lower and upper limits of the 95% confidence interval

Eq: Weights distributed equally over the whole sequence.

Be: 67% of the weight at the beginning.

End: 67% of the weight at the end.

L_2: L_2–norm method.

in Table 3; (b) 67% of the weight equally assigned to the first half of the longitudinal sequence, 33% equally assigned to the second half, and denoted by 'Be' in Table 3; (c) 33% of the weight equally assigned to the first half of the longitudinal sequence, 67% equally assigned to the second half, and denoted by 'End' in Table 3. The same notation as before was used for the confidence interval limits.

Note that, regardless of the set of weights used, Prolactine at day 3 always produced the best results, followed by Scotopic at day 3 and Mesopic Hi at day 3. Interestingly, we also observed some mild impact of the weights on the analysis. For instance, for Prolactine at day 3 the largest point estimate was obtained when most of the weights were assigned at the beginning of the sequence. Further, a closer look at Figure 2 corroborates that most of the effect appears at the beginning of the study and it fades away a bit towards the end. Similarly, for Scotopic at day 3, the largest point estimate was obtained when most of the weights were assigned at the beginning of

the sequence. Like before, here also Figure 2 confirms that most of the treatment effect is manifested for this biomarker at the beginning of the study. This clearly illustrates how the weights used in approach III could help to detect some interesting patterns in the longitudinal profiles even though they all lead to the same general conclusion.

7 Concluding Remarks

Biomarkers are playing an increasingly important role, not only in the study and development of new drugs and therapies, but also in the diagnostics of a medical condition or in improving our understanding of several medical conditions. The recent developments in genetics will likely further increase their utility and use in the near future. Even though considerable research has been done in recent years to study the potential of biological markers as surrogate endpoint; other possible uses have received less attention from a statistical point of view.

In the present work, we focused on the study and evaluation of different physiological variables as biomarkers for pharmacological activity. This type of studies are typically carried out following a cross-over design and include a relatively small group of patients. The use of a cross-over design in a longitudinal context will require special analysis considerations. In all cases, we decided to use a saturated linear model, guaranteeing the necessary flexibility to model the time evolution of a relatively large number of biomarkers. Further, we proposed three different approaches using multivariate and univariate techniques. Note that even though one could argue that the multivariate ellipsoid method is more powerful than the other alternatives, the L_2-norm and weighted procedures also offer a great flexibility to answer interesting scientific questions.

From a theoretical point of view, there is a rich potential arising from the possibility to vary the weighting scheme. For example, if varying the weighting scheme affects the results not at all or only very little, it could essentially imply that there is a uniform impact throughout time. In the reverse case, varying the weighting scheme could allow further exploration of the differential impact occurring across time. Indeed, in such a case, by choosing appropriate sets of weights, one could explore not only in which biomarker the treatment effect is expressed most but also in which biomarkers the treatment effect acts in a time-specific way, and whether such action is seen, for example, earlier or later in the sequence. Our simulations showed a negligible impact of the weighting scheme on the ranking of the biomarkers. However, the weights seem to have an effect on the power of the different procedures. It is clear that more research will be needed before a more complete idea about the potential and limitations of different sets of weights can emerge.

The three methods introduced in this work are clearly model dependent. More complicated models may be required to handle more complex designs, or to handle a period effect. However, the general ideas underneath the three approaches are fully general and could, in principle, be applied as well with more complex models after some adjustment. For instance, in the ellipsoid approach, the vector of parameters characterizing the treatment effect evolution over time may have different dimension and interpretation, depending on the specific modeling framework used for analysis. Nevertheless, the corresponding maximum likelihood estimator will still be asymptotically normal and Theorem **??** will still be valid.

Finally, note that biomarkers can serve a variety of purposes. In the study under consideration here, the main question was to find out in which biomarker the effect of treatment is seen most clearly, in other words, is expressed most. In other settings, there may be the desire to eventually use a biomarker as a surrogate marker or

surrogate endpoint; in such a scenario, one would like to predict the treatment effect on a so-called true endpoint, using the treatment's effect on the surrogate. This requires appropriate and somewhat different technology. An important contribution to this effect has been made by Burzykowski and Buyse (2006), through their so-called *surrogate threshold effect*.

Acknowledgment

Ariel Alonso and Geert Molenberghs gratefully acknowledge financial support from the Interuniversity Attraction Pole Research Network P6/03 of the Belgian Government (Belgian Science Policy). All authors are grateful to Eli Lilly & Company for graciously providing the data.

8 References

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapy*, **69**, 89–95.

Burzykowski, T. and Buyse, M. (2006). Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, **5**, 173–186.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer.

Jones, B. and Kenward, G.M. (2003). *Design and Analysis of Cross-over Trials*. London: Chapman & Hall/CRC.

Lesko, L.J.S. & Atkinson, A.J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annals Review of Pharmacology and Toxicology*, **41**, 347-366.

Suico, J., Renard, D., Carter, M.K., Diringer, K., Cantrell, A., Chavers, J., and Chalon, S. (2005). Pharmacodynamic effects of opiate receptor antagonism: an agonist challenge study. *Clinical Pharmacology & Therapeutics*, **77**, 28-28.

Temple R.J. (1995). A regulatory authority's opinion about surrogate endpoints. In: W.S. Nimmo and G.T. Tucker G.T. (Eds.), *Clinical Measurement in Drug Evaluation*, New York: John Wiley & Sons, pp. 3–22.

Appendix

Proof of Theorem 1

To simplify the notation we shall denote $\boldsymbol{\Delta}_T = \boldsymbol{x}$, and $\widehat{\boldsymbol{\Delta}}_T = \boldsymbol{d}$ and $\Sigma = \widehat{\Sigma}_{\Delta T}$. We can then rewrite the previous expressions as:

$$r = \min \parallel \boldsymbol{x} \parallel^2 = \boldsymbol{x}'\boldsymbol{x}$$

$$\text{st:} \quad (\boldsymbol{x}-\boldsymbol{d})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) = C(\alpha).$$

Using Lagrange's method, our problem is reduced to minimizing the following function: $F(\boldsymbol{x},\lambda) = \boldsymbol{x}'\boldsymbol{x} + \lambda(\boldsymbol{x}-\boldsymbol{d})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) - \lambda C(\alpha)$. Equivalently, we have to solve the simultaneous equations:

$$\frac{\partial F}{\partial \boldsymbol{x}} = 2\boldsymbol{x} + 2\lambda\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) = 0, \tag{10}$$

$$\frac{\partial F}{\partial \lambda} = (\boldsymbol{x}-\boldsymbol{d})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) - C(\alpha) = 0. \tag{11}$$

It is not difficult to show that (10) leads to $\boldsymbol{x} = \lambda(\Sigma+\lambda I)^{-1}\boldsymbol{d}$. Additionally, we have that there exist an orthogonal matrix P so that $\Sigma = P'D_0P$ with $P'P = PP' = I$. D_0 is a diagonal matrix, i.e., $D_0 = (\alpha_i)_{ii}$, where α_i is the i^{th} eigenvalue of Σ. Using this orthogonal decomposition we see that $\boldsymbol{x} = \lambda P'(D_0+\lambda I)^{-1}P\boldsymbol{d}$.

If we now denote $D_1(\lambda) = \lambda(D_0+\lambda I)^{-1} = \text{diag}\left(\frac{\lambda}{\alpha_i+\lambda}\right)$, then $\boldsymbol{x} = P'D_1(\lambda)P\boldsymbol{d}$. Combining this last expressions for $\boldsymbol{x}$ with (11) we obtain

$$(\boldsymbol{x}-\boldsymbol{d})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) = \boldsymbol{q}'\left[D_3(\lambda) - 2D_2(\lambda) + D_0^{-1}\right]\boldsymbol{q},$$

where $D_2(\lambda) = \text{diag}\left(\frac{\lambda}{\alpha_i(\alpha_i+\lambda)}\right)$, $D_3(\lambda) = \text{diag}\left(\frac{\lambda^2}{\alpha_i(\alpha_i+\lambda)^2}\right)$, and $\boldsymbol{q} = P\boldsymbol{d}$.

The matrix of the previous quadratic form is symmetric with diagonal elements equal to $D_3(\lambda) - 2D_2(\lambda) + D_0^{-1} = \text{diag}\left(\frac{\alpha_i}{(\alpha_i+\lambda)^2}\right)$. If $\boldsymbol{q}' = (q_1, q_2, \ldots, q_m)$ then

$$(\boldsymbol{x}-\boldsymbol{d})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{d}) = \sum \frac{\alpha_i q_i^2}{(\alpha_i+\lambda)^2}, \tag{12}$$

The previous expression clearly illustrates that (11) is equivalent to the equation (4) defined in theorem 1

$$\sum \frac{\alpha_i q_i^2}{(\alpha_i + \lambda)^2} = c(\alpha). \tag{13}$$

Using (13) we calculate λ, and finally we just have to calculate the distance

$$r \quad = \quad \boldsymbol{x}'\boldsymbol{x} = \boldsymbol{q}' D_1(\lambda)^2 \boldsymbol{q} = \sum \left(\frac{\lambda q_i}{\alpha_i + \lambda} \right)^2 .$$