Investigating Association Between Behavior, Corticosterone, Heart Rate, and Blood Pressure in Rats Using Surrogate Marker Evaluation Methodology

TILAHUN ESHETE, Abel; MARINGWA, John; GEYS, Helena; ALONSO ABAD, Ariel; Raeymaekers, Leen; MOLENBERGHS, Geert; Van Den Kieboom, Gerd; Drinkenburg, Pim & BIJNENS, Luc (2009) Investigating Association Between Behavior, Corticosterone, Heart Rate, and Blood Pressure in Rats Using Surrogate Marker Evaluation Methodology. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 19(1). p. 133-149.

DOI: 10.1080/10543400802527924
Handle: http://hdl.handle.net/1942/9253

Investigating Association Between Behavior, Corticosterone, Heart Rate, and Blood Pressure in Rats Using Surrogate Marker Evaluation Methodology

Abel Tilahun[1], **John T. Maringwa**[1], **Helena Geys**[2], **Ariel Alonso**[1], **Leen Raeymaekers**[2], **Geert Molenberghs**[1], **Gerd Van Den Kieboom**[2], **Pim Drinkenburg**[2], **Luc Bijnens**[2]

[1]Center for Statistics, Hasselt University, Diepenbeek, Belgium

[2] Johnson and Johnson Pharmaceutical Research and Development, a division of Janssen Pharmaceutica, Beerse, Belgium

Abstract

The drug development process involves identifying a compound and assessing its merit through rigorous pre-clinical and clinical trials. The pre-clinical stage is designed to assess the chemical properties of the new drug, as well as to determine the steps for synthesis and purification. In this stage of drug development, circumstances might dictate the use of alternative endpoints than the originally anticipated clinically relevant endpoint. In this regard, identification and evaluation of surrogate endpoints is of paramount importance. The validation methods enable to quantify degrees of association between the clinically relevant endpoint, also termed the true endpoint, and the alternative, surrogate endpoint. In this paper, we adapt the surrogate marker evaluation methodology of Alonso *et al.* (2003, 2006), developed for the case of two longitudinal outcomes, to the situation where either a longitudinal surrogate and cross sectional true endpoint is recorded, or vice versa. The work is motivated by a preclinical experiment conducted to asses association between Corticosterone (CORT), heart rate, and blood pressure in rats, the data from which are then subjected to analysis. It was found that there is a weak relationship between CORT and behavior, and between CORT on the one hand and heart rate and blood pressure on the other hand, but a reasonably high degree of association was registered between heart rate and behavior.

Some Key Words: fractional polynomial; R^2_Λ; spline; surrogate endpoint, true endpoint; variance reduction factor.

1 Introduction

The first step in the process of drug development is identifying promising compounds. Once a compound has been isolated for further scrutiny, it enters a rigorous testing and evaluation stage, the so-called pre-clinical phase. This stage is designed to assess the chemical properties of the new drug as well as to determine the steps for synthesis and purification. In this stage, the toxicological and pharmacological effects of the drug are evaluated through in-vitro and in-vivo animal testing. There might be a variety of reasons hindering undertaking these tests directly on the clinically relevant outcome, even when the studies involve animals, necessitating the use of biomarkers.

Several challenges are encountered in the identification of biomarkers, including: understanding the role of a specific biomarker to a clinically relevant problem; developing either an indirect or a direct readout of physiologic state; determining the comparable pathways between animal models and humans; and finally embedding the biomarker into a robust assay and subsequent validation and approval of the assay in clinical applications (Pien *e*t al., 2005). Several attempts, from both a biological and a statistical angle, have been made to circumvent these challenges (Burzykowski, Molenberghs, and Buyse, 2005). Focusing on the statistical problem of identifying and validating a biomarker, statistical expertise, in particular paradigms designed to validate surrogate markers, might be handy tools to quantify the degree of association between the biomarker and the clinically relevant outcome.

In surrogate marker evaluation, two possible sources of evidence can be sought to validate a biomarker. The first is situated at the individual patient level and is concerned with the biological pathway from the surrogate to the true endpoint. The second possible source of evidence comes from the trial level, and quantifies the association between the treatment effects on the marker and clinical endpoint (Burzykowski, Molenberghs, and Buyse, 2005). The purpose of this paper is to adapt existing surrogate marker validation methodology to quantifying the degree of association between behavior, as measured by alertness, corticosterone levels, and telemetry measures such as heart rate and blood pressure of rats, with emphasis given to the prediction of one of the outcomes given the other in a single trial setting. Note that, if there is an interest in the trial-level surrogacy, there is then need for repetition of the experiment, for example at different centers and/or by different investigators, or even through the conduct of a sequence of altogether different trials.

The rest of this paper is organized as follows. The motivating study is introduced in Section 2, followed by a description of flexible models for longitudinal data in Section 3. A review of the basic, single-trial approaches for continuous outcomes is offered in Section 4.1. Section 4.2 studies the so-called variance reduction factor (VRF) approach for combined longitudinal and cross-sectional endpoints. Section 4.3 is concerned with the so-called likelihood reduction factor (R^2_Λ) approach. Section 5 is dedicated to the case study's analysis.

2 Motivating Case Study

The data come from a preclinical rat experiment on a compound under development for stress-related disorders. The objective of the experiment was to identify the effect of the compound on stress hormones and a series of physiological variables. In the experiment, stress is induced by forcing a rat to swim for 15 minutes in a bath of 20 cm high lukewarm water of 25 degrees Celsius, according to a protocol as described by De Groote and Linthorst (2007). The experiment was designed according to a latin square crossover design with 4 periods and 4 treatment groups (vehicle without stress, vehicle with stress, compound without stress, compound with stress). Forty-five minutes after randomization, the rats were injected with either a vehicle or the compound under consideration. Ten minutes later, half of the rats injected with the vehicle and half of the rats injected with the compound, were subjected to so-called "swim stress", also depending on group membership. For all eight animals, measurements were analyzed in order to quantify their stress level. Telemetry measurements (such as heart rate and blood pressure) were recorded continuously and averaged every 5 minutes. Seventeen blood samples were taken in a fully automated way, leaving the animals completely undisturbed and following a well-defined scheme to sample blood plasma from which corticosterone (CORT) was later extracted and quantified . And finally, rats were also screened for their behavior in a 10 minutes interval by means of a video monitor. For each rat, the percentage of time it has been active is thus determined. The recording of behavior was done twice: a first time at 25 minutes after injection and a second time at 50 minutes after the end of the swim stress.

3 Flexible Linear Mixed Modeling

3.1 Longitudinal Data Analysis

The data considered here include both cross-sectional and longitudinal outcomes. Let us first give a brief introduction to the analysis of longitudinal data. Since we are in the framework of continuous longitudinal data, modeling can be done by way of a linear mixed model. The general linear mixed-effects model can be represented as (Verbeke and Molenberghs, 2000):

$$\begin{cases} \boldsymbol{Y}_j = X_j\boldsymbol{\beta}_j + Z_j\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j \\ \boldsymbol{b}_j \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\varepsilon}_j \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_j), \quad \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \quad \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \text{ are independent,} \end{cases} \tag{1}$$

where $\boldsymbol{Y}_j$ $(j = 1, \ldots, n)$ is the m_i-dimensional response vector of measurements for dog j, X_j and Z_j are $m_j \times p$- and $m_j \times q$-dimensional matrices of known covariates (e.g., time), respectively, $\boldsymbol{\beta}_j$ is a p-dimensional vector of fixed effects, $\boldsymbol{b}_j$ is q-dimensional dog-specific vector of random effects and $\boldsymbol{\varepsilon}_j$ is an m_j-dimensional vector of residuals. The matrix $\boldsymbol{G}$ is a general $q \times q$ covariance matrix and $\boldsymbol{\Sigma}_j$ is an $m_j \times m_j$ covariance matrix. Often, $\boldsymbol{\Sigma}_j$ is assumed to equal $\sigma_\varepsilon^2 \boldsymbol{I}_{m_j}$, resulting in the so-called conditional independence model. Note that when the response is cross-sectional, the general model reverts to the usual regression model wherein subject-specific effects are dropped.

The evolution over time can be captured by specifying parametric functions, such as, for example, linear, quadratic or even higher-order polynomials in the vector X_j. These effects may well be included in the random-effects vector Z_j as well. However, it is not difficult to imagine cases where obtaining a suitable parametric form adequately describing the mean is a challenge. Although our primary goal is to quantify the association between CORT, heart rate, and blood pressure via surrogate marker validation methods, proper modeling of the mean evolution in time is necessary. One can get rid of the need to specify a parametric model through use of flexible modeling techniques, an issue taken up further in the following section.

3.2 Flexible modeling techniques

Postulating a parametric function to model the mean evolution may be difficult and/or restrictive, as is clear from Figure 1. An appealing alternative is to model the time evolution using some flexible smooth function. In this section, we briefly discuss linear mixed models to model longitudinal data (Verbeke and Molenberghs, 2000) with the time trend determined by some flexible smooth function in the form of either penalized smoothing splines (Eilers and Marx, 1996; Verbyla *et al.*, 1999; Ruppert, Wand, and Carroll, 2003) or fractional polynomials (Royston and Altman, 1994).

3.2.1 Penalized Smoothing Splines

Use of penalized splines results in a semi-parametric smooth function, the term 'semi-parametric' here referring to the feature that the model combines parametric and non-parametric aspects. We provide a brief description of the model as is usually encountered with longitudinal data.

Let Y_{jk} denote the response taken from subject j at time t_{jk} $(k = 1, \ldots, K)$. The model of interest can be expressed as: $Y_{jk} = f(t_{jk}) + \varepsilon_{jk}$, for a smooth function $f(\cdot)$. Restricting focus to the truncated lines basis, which is simple in formulation and performs adequately in many circumstances (Ngo and Wand, 2004), the penalized-spline representation can be written as:

$$Y_{jk} = \beta_0 + \beta_1 t_{jk} + \sum_{q=1}^{Q} b_q (t_{jk} - \kappa_q)_+ + \varepsilon_{jk}, \tag{2}$$

where $\kappa_1, \ldots, \kappa_Q$ are a set of distinct knots in the range of t_{jk}, $t_+ = \max(0, t)$, and $b_q \sim N(0, \sigma_b^2)$. The knot points are selected as equally spaced quantiles of time (Ruppert *et al.*, 2003). For ease of development, we adopt the following matrix notation. Let

$$\boldsymbol{Y}_j = \left[\begin{array}{c} y_{jk} \end{array} \right]_{1 \le j \le n, 1 \le k \le K}, \qquad \boldsymbol{X}_j = \left[\begin{array}{cc} 1 & t_{jk} \end{array} \right]_{1 \le j \le n, 1 \le k \le K}, \qquad \boldsymbol{\beta} = \left[\begin{array}{cc} \beta_0 & \beta_1 \end{array} \right]'.$$

Further, define:

$$\boldsymbol{Z}_j = \left[\begin{array}{c} (t_{jk} - \kappa_k)_+ \end{array} \right]_{1 \le j \le n,\ 1 \le k \le K,\ 1 \le \kappa \le Q}, \quad \boldsymbol{b} = \left[\begin{array}{c} b_1, \ldots, b_Q \end{array} \right]', \qquad \boldsymbol{\varepsilon}_j = \left[\begin{array}{c} \varepsilon_{11}, \ldots, \varepsilon_{nK} \end{array} \right]'.$$

Using this notation, a stacked version of (2) becomes $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$. The correspondence between the penalized spline smoother and the optimal predictor in a mixed model framework is a key feature in fitting the models. This connection offers the opportunity of using ordinary software packages for mixed models, such as, for example, SPlus, SAS, or R. Here, we use the MIXED procedure in SAS.

Fitting penalized splines by the linear mixed model approach has some appealing advantages, such as automatic determination of the smoothing parameter, a unified framework for inference, and the flexibility with which the models can be extended (Faes *et al.*, 2006). Specifically, note that extending (2) to the cross-over setting simply involves addition of fixed effects, typically, treatment group, period, and carry-over effects, together with interactions of interest. The theory relating to cross-over designs is well established, whether for the simplest setting of two treatments and two periods, or for higher-order designs (Senn, 1993; Jones and Kenward, 2003).

3.2.2 Fractional Polynomials

As an alternative to capturing the time trend as mentioned in Section 3.2.1, the so-called fractional polynomial approach may be used. Fractional polynomials provide an extension to classical polynomials allowing for non-integer powers to the time covariate, thereby adding greater flexibility in capturing rather complex non-linear relationships. A brief description of fractional polynomials is given below.

Let $\boldsymbol{t} = (t_{j1}, \ldots, t_{jK})$ denote the set of time points pertaining to subject j. Royston and Altman (1994) define a fractional polynomial of degree m by

$$\phi_m(\boldsymbol{t}; \boldsymbol{\beta}, \boldsymbol{p}) = \sum_{r=0}^{m} \beta_r H_r(\boldsymbol{t}), \tag{3}$$

where m is a positive integer and $\boldsymbol{p} = (p_1, \ldots, p_m)$ is a real-valued set of powers such that $p_1 \leq \cdots \leq p_m$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ are real-valued coefficients. For $r = 0$, $H_0(\boldsymbol{t}) = 1, p_0 = 0$, and for $r = 1, \ldots, m$:

$$H_r(\boldsymbol{t}) = \begin{cases} \boldsymbol{t}^{p_r} & \text{if } p_r \neq p_{r-1}, \\ H_{r-1}(\boldsymbol{t}) \ln(\boldsymbol{t}) & \text{if } p_r = p_{r-1}. \end{cases}$$

As mentioned in Royston and Altman (1994), polynomials of a degree higher than 2 or 3 are rarely encountered in practice. The best power transformation is frequently found among the members of the list $\{-2, -1, -0.5, 0, 0.5, 1, \ldots, \max(3, m)\}$. While it is possible to incorporate other powers, there is a danger coming with including large negative powers, in the sense that individual extreme observations will influence the fit too much (Royston, Parmar, and Qian 2003).

Note that the fractional polynomial model has been defined in its generic form and in analogy with penalized-splines models; extension to include covariates other than time is possible. In such a situation, an extension of (3) may be obtained through adding the fixed effects for treatment, period, and carry-over, together with relevant interactions.

4 Validation Methods

In this section, we give a concise description of the various methods used in validating a surrogate endpoint, with emphasis on the individual level.

4.1 Review of the Single Trial-based Validation Methods for Continuous Outcomes

Several methods have been suggested for the formal evaluation of surrogate markers. Some of these methods are based on a single trial while others, which are gaining momentum in the present day, are based on meta-analytic concepts. The first formal approach to evaluate markers is attributed to Prentice

(1989), who has given a definition of surrogate endpoints, followed by a series of operational criteria to check whether the definition is fulfilled.

Freedman, Graubard, and Schatzkin (1992) have supplemented the hypothesis-testing-based criteria, which necessarily depend on the power of the test performed, with a quantity to be estimated. They suggested the use of the so-called *proportion of treatment effect explained* (PTE) by the surrogate as an alternative means of validation. The PTE faces serious drawbacks, against the background of which Buyse and Molenberghs (1998) have suggested the use of another quantity, the *relative effect* (RE), defined as the ratio of the treatment effect on the true endpoint to that on the surrogate endpoint. In turn, the RE is open to severe criticism as well. First, the RE's confidence intervals, like the ones for PTE, tend to be wide. While this could in principle be overcome, there is a second, more severe problem in the sense that the RE is useful for prediction of the true treatment effect from the surrogate treatment effect only when the relationship between both is multiplicative. This may be rightfully viewed as restrictive and, in any case, cannot be verified from a single trial.

Switching to the individual patient or experimental animal level, the need might arise to quantify the association between the surrogate and the true endpoint after adjustment for the treatment effect. To this end, Buyse and Molenberghs (1998) suggested the use of the adjusted association. Suppose we have a single trial, let S_j and T_j be the surrogate and true endpoint, respectively, and let Z_j be a binary treatment group indicator. To compute the adjusted association, consider the following pair of models:

$$\begin{aligned} S_j &= \mu_S + \alpha Z_j + \varepsilon_{S_j}, \\ T_j &= \mu_T + \beta Z_j + \varepsilon_{T_j}, \end{aligned}$$

where the error terms have a joint zero-mean normal distribution with variance-covariance matrix:

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc} \sigma_{SS} & \sigma_{ST} \\ \sigma_{TS} & \sigma_{TT} \end{array} \right).$$

Then, the *adjusted association*, denoted R^2 can be computed as:

$$R^2 = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}.$$

Note that the individual-level surrogacy is meant to measure the degree of correlation between the two endpoints after correcting for treatment and other possible effects. One way to correct for treatment effect is by including a treatment variable in the model, thus justifying inclusion of treatment effect in the model.

4.2 Variance Reduction Factor

In this section, we review the variance reduction factor, suggested by Alonso *et al.* (2003) for the case of two repeatedly measured outcomes, where after we show how this method can be adapted to the situation where one of the two outcomes is cross-sectional. Let us assume that there are n subjects enrolled for a particular study and further suppose that t_{jk} is the time at which the kth measurement of the jth subject is taken. Let T_{jk} and S_{jk} be the true and the surrogate endpoints, respectively, and let Z_j be a binary treatment indicator. Now, consider the following joint model for the true and surrogate endpoints:

$$\begin{aligned} T_{jk} &= \mu_T + \alpha Z_j + f(t_{jk}) + \varepsilon_{Tjk}, \\ S_{jk} &= \mu_S + \beta Z_j + f(t_{jk}) + \varepsilon_{Sjk}, \end{aligned} \tag{4}$$

where (μ_T, μ_S, α,β) are intercepts and treatment effects on the true and surrogate endpoints, respectively, $f(t_{jk})$ is a flexible function in time which can be modeled as fractional polynomial, penalized spline, or any flexible function in time. In principle, it is possible for the two endpoints to depend on time through different functions, in which case we will have $f_T(t_{jk})$ and $f_S(t_{jk})$ for the true and surrogate endpoint respectively. However, without loss of generality, let us assume that both depend on time through the same function. The error terms (ε_{Tjk}, ε_{Sjk}) are assumed to follow a zero-mean normal distribution with patterned variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}, \tag{5}$$

with obvious notation.

In this setting, Alonso *et al.* (2003) proposed to quantify the individual-level surrogacy using the so-called *variance reduction factor*, which is defined as

$$VRF = \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})}, \tag{6}$$

where $\Sigma_{T|S}$ denotes the conditional variance-covariance matrix of T_{jk} given S_{jk}, i.e., $\Sigma_{T|S} = \Sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Furthermore, these authors have shown that the VRF satisfies a set of properties that makes it practically applicable: (i) VRF ranges between zero and one; (ii) $VRF = 0$ if and only if the true and the surrogate endpoints are independent; (iii) $VRF = 1$ if and only if there exists a deterministic relationship between the true and surrogate endpoint; and (iv) $VRF = R^2$ in the cross-sectional setting.

Note that, at the individual level, interest lies in the prediction of the true endpoint given the surrogate endpoint. In this regard, property (ii) shows that if the VRF equals zero, then no sensible prediction is possible, whereas a perfect prediction is attained if VRF equals one, as indicated by property (iii). Property (iv) establishes the link between this approach and the one suggested by Buyse *et al.* (2000) for univariate outcomes.

Let us now turn to the question as to how this approach can be used when one of the two endpoints is cross-sectional. Assume we have K measurements per subject for the longitudinal outcome.

4.2.1 Case 1: A Longitudinal Surrogate for a Cross-sectional True Endpoint

Let us assume that the surrogate endpoint is repeatedly measured over time with K repeated measures and that the true endpoint is cross-sectional. Model (4) takes the form:

$$\begin{aligned} T_j &= \mu_T^* + \alpha^* Z_j + \varepsilon_{Tj}, \\ S_{jk} &= \mu_S^* + \beta^* Z_j + f(t_{jk}) + \varepsilon_{Sjk}, \end{aligned} \tag{7}$$

and the error terms $(\varepsilon_{Tj}, \varepsilon_{S_{jk}})$ are assumed to follow a zero-mean normal with variance-covariance matrix Σ, which in this setting takes the form

$$\Sigma = \begin{pmatrix} \sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}. \tag{8}$$

Here, σ_{TT} denotes the variance of the true endpoint, Σ_{TS} is a $(1 \times K)$ vector containing the covariances between the true endpoint and the surrogate endpoint at different time points, and Σ_{SS} is a $(K \times K)$ variance-covariance matrix associated with the longitudinal surrogate endpoint. Then, the VRF_{indiv} for longitudinal surrogate and a cross-sectional true endpoint denoted by VRF_{ST}^{LC}, with a subscript 'L' ('C') reminiscent of 'longitudinal' ('cross-sectional'), can be computed as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{T|S})}{\text{tr}(\sigma_{TT})}, \tag{9}$$

where $\sigma_{T|S}$ denotes the conditional variance of T given S: $\sigma_{T|S} = \sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Using this expression, (9) can be re-written as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\sigma_{TT})}. \tag{10}$$

Note that all matrices involved in the computation of VRF_{ST}^{LC} are of dimension (1×1) and hence the trace reduces to the corresponding scalar, offering the opportunity to simplify (10):

$$VRF_{ST}^{LC} = \frac{\sigma_{TT} - \sigma_{TT} + \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} = \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}. \tag{11}$$

Notice that $VRF_{ST}^{LC} = 0$ if and only if $\Sigma_{ST} = 0$, i.e., if and only if when S and T are independent.

Intuitively, (11) quantifies how much of the total variability of the true endpoint is explained by the surrogate endpoint, after adjusting for treatment effects and repeated measures of the surrogate endpoint.

4.2.2 Case 2: A Cross-sectional Surrogate for a Longitudinal True Endpoint

Next, let us consider a role reversal, such that the true endpoint is repeatedly measured over time with K repeated measures, whilst having the surrogate endpoint in cross-sectional form. Model (4) becomes:

$$\begin{array}{rcl} T_{jk} & = & \mu_T^* + \beta^* Z_j + f(t_{jk}) + \varepsilon_{Tjk}, \\ S_j & = & \mu_S^* + \alpha^* Z_j + \varepsilon_{Sj}. \end{array} \tag{12}$$

The error terms $(\varepsilon_{Tjk},\varepsilon_{S_j})$ are zero-mean normally distributed with variance-covariance matrix:

$$\Sigma = \left(\begin{array}{cc} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \sigma_{SS} \end{array} \right), \tag{13}$$

Now, the VRF_{indiv} for this case is

$$\begin{array}{rcl} VRF_{ST}^{CL} & = & \dfrac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})} \\ & = & \dfrac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{TT} - \Sigma_{TS}\sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\Sigma_{TT})} \\ & = & \dfrac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{TT}) + \text{tr}(\Sigma_{TS}\sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\Sigma_{TT})} \\ & = & \dfrac{\text{tr}(\Sigma_{TS}\Sigma_{ST})}{\sigma_{SS}.\text{tr}(\Sigma_{TT})}. \end{array} \tag{14}$$

From (11) and (14), it is clear that there is asymmetry in the VRF calculations. Results differ depending on which of the two endpoints is the cross-sectional one. This is in line with our expectations. In the case of a longitudinal true endpoint, the VRF measures the ability of the cross-sectional endpoint to predict the longitudinal outcome at each time point, whereas when the longitudinal sequence is treated as surrogate endpoint, the VRF measures the adequacy of the longitudinal sequence to predict the cross-sectional outcome. It is therefore imperative to determine in advance which of the two outcomes is treated as true when applying this procedure to quantify association. Either way, a VRF value close to one indicates that the surrogate is a 'good' predictor of the true endpoint at the individual level, while values close to zero indicate 'poor' prediction. In any case however, the values of the VRF have to be complemented with expert opinion before passing judgment on the adequacy of the surrogate to predict the true endpoint.

4.3 The Measure R^2_Λ

As can be seen from (6), the VRF summarizes the variability of the two endpoints using the trace of the corresponding variance-covariance matrices. In multivariate analysis, there is no unique way of defining a generalized variance, the trace is one of the classical ways of doing so, while another common definition uses the determinant. Interestingly, using the trace or the determinant to summarize the variability of the endpoints has important ramifications for analysis and leads to two totally separate measures with different interpretations. To this end, Alonso *et al.* (2006) have suggested another measure, the so-called R^2_Λ, which uses this alternative definition of the generalized variance. Like the VRF, this measure can be derived based on Model (4), as follows:

$$R^2_\Lambda = 1 - \frac{|\Sigma\,|}{|\Sigma_{TT}\,| \cdot |\Sigma_{SS}\,|}. \tag{15}$$

The authors have shown that this measure enjoys desirable properties: (i) R^2_Λ is symmetric and invariant with respect to linear bijective transformations; (ii) R^2_Λ ranges between zero and one; (iii) $R^2_\Lambda = 0$ if and only if the error terms are independent; (iv) $R^2_\Lambda = 1$ if and only if there exist a and b so that $a^T\varepsilon_{S_{jk}}$ $= b^T\varepsilon_{T_{jk}}$ with probability one; and (v) $R^2_\Lambda = R^2$ in the cross-sectional setting.

All of these properties, except the fourth property are shared with the VRF. The fourth property, however, differs in important ways from the VRF. Indeed, whereas the VRF takes the value 1 when there is a deterministic relationship between both endpoints, R^2_Λ is 1 whenever there is a deterministic relationship between two linear combinations of both endpoints, allowing us to uncover strong association in cases where the VRF might fail to do so. This is not a disadvantage of one or the other proposal, but rather underscores them focusing on different aspects. The expression for R^2_Λ clearly shows that, unlike the VRF, this measure treats both endpoints symmetrically. To clarify this further, let us first consider the surrogate to be longitudinal and the true endpoint cross-sectional, and thereafter reverse the roles.

4.3.1 Case 1: A Longitudinal Surrogate for a Cross-sectional True Endpoint

Consider Model (7) and the corresponding variance-covariance matrix (8). The R^2_Λ for a longitudinal surrogate and a cross-sectional endpoint is given by

$$R^{2,LC}_{\Lambda,ST} = 1 - \frac{|\Sigma\,|}{|\sigma_{TT}\,| \cdot |\Sigma_{SS}\,|}, \tag{16}$$

where σ_{TT}, Σ_{SS}, and Σ are as defined in (8). Note that

$$|\Sigma| = |\Sigma_{SS}| \cdot |\Sigma_{T|S}| = |\Sigma_{SS}| \cdot |\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}|$$

and, substituting this in (16), we obtain

$$\begin{aligned} R^{2,LC}_{\Lambda_{ST}} &= 1 - \frac{|\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}\,|}{|\sigma_{TT}\,|} \\ &= 1 - \frac{\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} \\ &= \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}, \end{aligned} \tag{17}$$

since all matrices involved are of dimension one.

4.3.2 Case 2: A Cross-sectional Surrogate for a Longitudinal True Endpoint

Consider model (12). The R^2_Λ for a longitudinal true and a cross-sectional surrogate endpoint is

$$\begin{aligned} R^{2,CL}_{\Lambda,ST} &= 1 - \frac{|\Sigma\,|}{|\Sigma_{TT}\,| \cdot |\sigma_{SS}\,|} \\ &= 1 - \frac{|\Sigma_{TT}\,| \cdot |\sigma_{SS} - \Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}\,|}{|\Sigma_{TT}\,| \cdot |\sigma_{SS}\,|} \\ &= 1 - \frac{|\sigma_{SS} - \Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}\,|}{|\sigma_{SS}\,|} \\ &= 1 - \frac{\sigma_{SS} - \Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}}{\sigma_{SS}} \\ &= \frac{\Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}}{\sigma_{SS}}. \end{aligned} \tag{18}$$

Comparing (17) with (18) establishes that $R^{2,LC}_{\Lambda,ST} = R^{2,CL}_{\Lambda,ST}$. In the first case, we used σ_{TT} and Σ_{SS} as component variances, of scalar and matrix type, respectively. These roles are reversed in the current, second case. Nevertheless, we obtain the same final expression for R_{Λ}^{2} as is, of course, entirely in line with the original, symmetric definition (15) of the quantity.

Furthermore, note that R_{Λ}^{2} and VRF are equal when the surrogate is longitudinal and the true endpoint cross-sectional. This implies that, only the VRF with the surrogate cross-sectional and the true endpoint longitudinal will be different from all of the others, that than coincide. This again highlights the feature that, for a longitudinal true endpoint, the VRF studies prediction of the entire sequence, while the R_{Λ}^{2} assesses how well an optimal linear combination of the true endpoint profile can be predicted. Both may be useful, but definitely are different. Moreover, one would expect the VRF to be well below the R_{Λ}^{2} in many applications, since prediction of an entire longitudinal sequence from a cross-sectional quantity is a tall order, whereas it might well be feasible to predict a particular linear combination. The choice between the two measures lies in the objective to be attained. If the objective is to measure the strength of the surrogate to predict the entire sequence of the true endpoint, then VRF will be an ideal choice. However, when this seems an attainable goal or when we are rather interested in predicting some linear combination of the true endpoint, then we can resort to R_{Λ}^{2}. Note that standard error of the estimates can be calculated using either a delta method or bootstrap (Efron, Bradley and Tibshirani, 1993). In this manuscript, standard errors are calculated using bootstrap methods.

5 Application to the Case Study

The treatment is referred to as "the compound" due to confidentiality. The experiment was performed in the following manner. First, the eight rats were randomized to treatment and vehicle groups and then followed up, a time during which several responses were measured. After a fixed period, half of the rats from the treatment group and half of those from the vehicle group were subjected to stress. The follow-up then continued and some further measurements ere recorded. Thus, after stress, there were four groups: (1) "treatment alone;" (2) "treatment+stress;" (3) "vehicle alone;" and (4) "vehicle +stress." The researchers wished to assess the association between the different responses before and after stress was induced. Thus, the results for pre- and post-stress correspond to the associations measured between the different responses before and after the stress with the treatment variable (Z), having two possible values (1: active treatment, 0:vehicle) for pre-stress and having four different possible values after stress.

Figure 1 shows the group-specific mean profiles of CORT measurements, averaged over the four treatment periods. The plot depicts the average CORT values per treatment group at each time point, essentially showing how, on average, CORT values evolve over time in each treatment group. A detailed account on how these data were collected can be found in Section 2. The need for flexible modeling tools is apparent from Figure 1, since finding a suitable or rather an acceptable classical parametric model is not an easy task. Hence, as mentioned before, we discuss results emanating from an application of surrogate marker validation methodology in conjunction with flexible modeling techniques (spline and fractional-polynomial based), meant to appropriately capture trends over time. For purposes of comparison, an unstructured mean model or a full factorial structure for time is also considered. However, this approach often yields excessively large numbers of parameters, thereby rendering it less desirable.

The VRF and R^2_Λ approaches have been applied to the dataset introduced in Section 2. The variance-covariance matrices, based upon which the VRF and R^2_Λ are computed, are estimated using maximum likelihood. The variance-covariance matrices can assume general structures unless the data suggests otherwise. In such cases, simple covariance structures, such as auto-regressive or compound symmetry, might be considered. For the purpose of our application, a number of models with different variance-covariance structures has been fitted. The best model, here being an unstructured variance-covariance structure, can be chosen using a conventional likelihood ratio test and/or Akaike's Information Criterion. The results of the analysis for the association of telemetry and behavior as well as that of CORT and behavior are summarized in Table 1 with bootstrap standard errors and in Table 2 with asymptotic standard errors, respectively. We should like to point out that it is not a trivial task to derive a closed-form expression for the standard errors of VRF and R^2_Λ for the particular case we have considered. However, fortunately, Alonso *et al* (2006) have shown that the VRF and R^2_Λ are special cases of the so-called *Likelihood Reduction Factor*, which is based on the information-theory approach. These authors have derived an asymptotic solutions for LRF. Hence, by virtue of the relationship of these measures with the LRF, we have been able to provide asymptotic standard errors based on the information-theory approach. There are no general guidelines as to how large a VRF and R^2_Λ should be in order to be considered sufficiently large. However, since the VRF and R^2_Λ are R-square type measures, it might be possible to make some general remarks concerning the degree of association based on their magnitude.

Since such a degree of association arguably would vary from application to application, the final decision has to be made in consultation with the experts, regardless their value. Having this in mind, from the results for the pre- and post-stress, we might infer that there is a rather weak relationship between behavior and CORT. However, strong and moderate relationships were observed between heart rate and behavior, and between blood pressure and behavior, respectively. Recall that behavior is measured cross-sectionally while CORT, heart rate, and blood pressure are longitudinal outcomes.

In this regard, when the cross-sectional outcome was used as a possible surrogate for the longitudinal outcomes, the VRF produced very low values, as anticipated in the previous section. Indeed, it is very difficult to predict the subtleties and richness of a longitudinal sequence from a single, cross-sectional measure. We consider this a desirable feature of the VRF. The R^2_Λ on the other hand, states that, although still small for some of the endpoints, there is better hope to predict a particular linear combination of the longitudinal outcomes from the cross-sectional outcome. As such, VRF and R^2_Λ both provide useful but totally *different* pieces of information. When there is role reversal, that is, when the longitudinal outcomes were treated as a possible surrogates for the cross-sectional outcome, the VRF values coincided with the R^2_Λ. This underscores that the VRF does not treat both endpoints symmetrically. The R^2_Λ, however, stayed the same even when there was role reversal, as expected from its construction.

The higher VRF and R^2_Λ values obtained when the longitudinally measured heart rate and blood pressure were used as surrogate endpoints for the cross-sectionally measured behavior, establish the possibility of predicting behavior using some linear combination of the longitudinal sequence.

Zooming in on the association between telemetry and CORT, both longitudinal in nature, we learn that there is a very weak association, with a maximum $\widehat{R}^2_\Lambda = 0.2314$ and maximum $\widehat{VRF} = 0.0513$, between the three modeling approaches. This is an indication that there is a very limited overlap in information between both outcomes, inhibiting comfortable prediction of one from the other.

In conclusion, the analysis has revealed that the longitudinally measured CORT level offers limited opportunity for prediction of activity, which is measured by the degree of alertness expressed in terms of the percentage of minutes the rats have been awake. We learn that heart rate and blood pressure are weakly related to CORT but have a strong predictive ability for behavior. The results advice against the use of activity to predict the longitudinal CORT level, heart rate, and blood pressure at each time point. These findings, however, have to be complemented with expert opinion before the results are to be practically used.

6 Discussion

In this manuscript, we have adapted surrogate marker evaluation methods, originally designed to handle two repeated measures sequences, to the case of one cross-sectional and one longitudinal outcome, where either of these can be used as the surrogate. The methods have been applied to quantifying association between longitudinally measured CORT level, heart rate, and blood pressure, with cross-sectional behavior measured by the level of activity, expressed as the percentage of time experimental rats have been active after exposure to treatment followed by stress. The methods appear to work adequately for this particular mix of longitudinal and cross-sectional endpoints.

The various theoretical properties of the methods have manifested themselves in the results of the data analysis. In particular, it has been nicely confirmed that the VRF focuses on the prediction of a longitudinal sequence *as a whole* by a cross-sectional outcome, while R^2_Λ is concerned with the prediction of an *optimal linear combination* of the longitudinal outcome.

In the case of two longitudinal outcomes, the optimal linear combinations from the two outcomes are the first canonical variates. In the context of a longitudinal true and cross-sectional surrogate endpoint, the optimal linear combination could be the first principal component or any other summary measure of the longitudinal measurements, thereby maximally retaining information. Thus, optimality in this context refers to finding a linear combination that best summarizes the repeated measures.

The longitudinal outcomes were modeled using flexible modeling tools such as fractional polynomials, penalized splines, and a general unstructured mean where the time trend is not modeled but rather an analysis-of-variance type approach is followed. This offers the possibility of fitting different models and then selecting the best one according to some model selection tool such as, for example, Akaike's Information Criterion. It is, indeed, important to conduct proper modeling before moving into quantifying surrogacy, because the results may critically depend on the model's goodness-of-fit.

In all cases, VRF or R^2_Λ estimates close to one are indicative of 'good' surrogacy, with the reverse holding for values close to zero. Evidently, it is difficult to provide general advice as to how large is large enough. Arguably, the statistical evaluation of a surrogate can be an important component in the decision making process, but at least equally important is expert opinion coming in from pharmacological, biological, clinical, ethical, and health economy considerations.

All analysis performed in this paper can be conducted using statistical softwares such as SAS, SPlus, R, or any package that allows fitting of bivariate models. A SAS macro is available from the authors' web site `www.uhasselt.be/censtat`.

Acknowledgment

Financial support from the IAP research network #P6/03 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

References

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M.G. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measures. *Biometrical Journal*, **45**, 931–945.

Alonso, A., Molenberghs, G., Geys, H., and Buyse, M. (2006). A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine*, **25**, 205–211.

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

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics*, **1**, 49–67.

De Groote, L. and Linthorst, A.C. (2007). Exposure to novelty and forced swimming evoke stressor-dependent changes in extracellular GABA in the rat hippocampus. *Neuroscience*, **148**, 794–805.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Faes, C., Geys, H., Molenberghs, G., Aerts, M., Cadarso-Suarez, C., Acuña, C., and Cano, M. (2006). A flexible method to measure synchrony in neuronal firing. *Journal of American Statistical Association*, **101**, 000–000.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

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

Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1–56.

Pien, H.H., Fischman, A.J., Thrall, J.H., and Sorensen, A.G. (2005). Using imaging biomarkers to accelerate drug development and clinical trials. *Drug Disc Today*, **10**, 259–266.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Royston, P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.

Royston, P., Parmar, M.K.B., and Qian, W. (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine*, **22**, 2239–2256.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Senn, S. (1993). *Cross-over Trials in Medical Research*. Chichester: John Wiley.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.

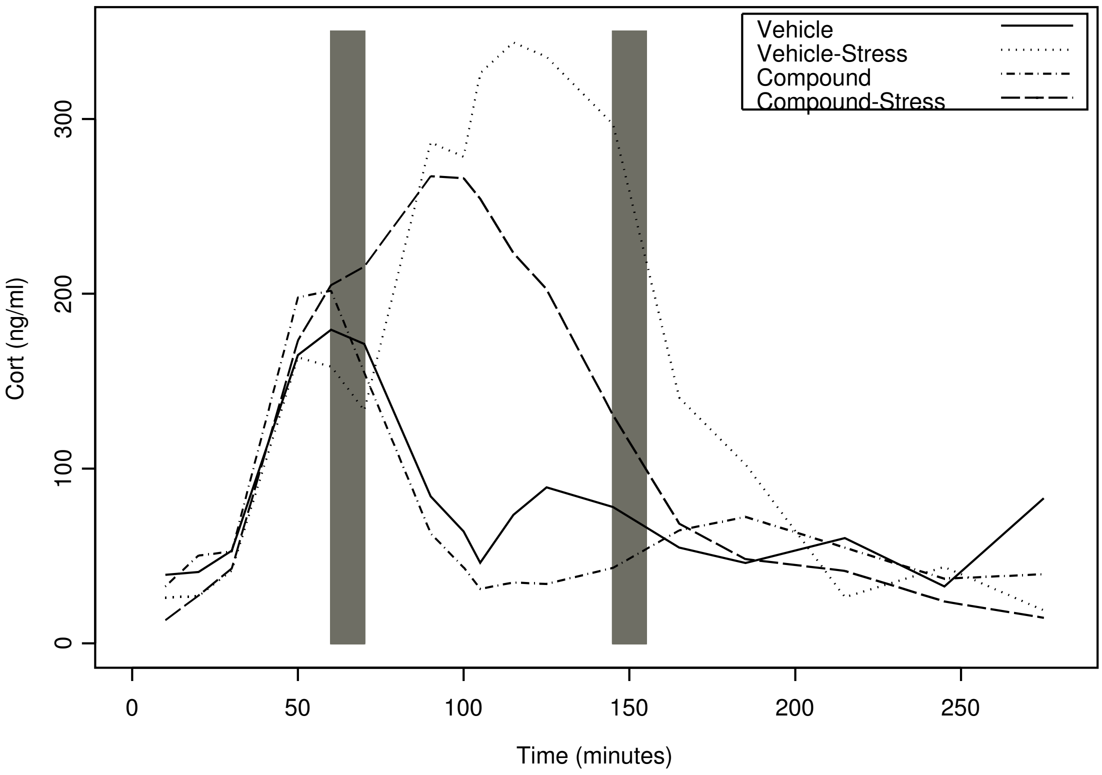

Figure 1: *Group-specific mean profiles of CORT values, averaged over different treatment periods. The shaded regions indicate the time windows in which activity was measured before and after the stress induction.*

Table 1: R^2_{indiv} *values(bootstrap standard errors) under pre-stress and post-stress conditions, for a variety of true and surrogate endpoints, using unstructured, fractional polynomial, and penalized splines models, and based on both* VRF *and* R^2_Λ.

endpoint		unstructured		fract. pol.		pen. splines	
true	surrogate	VRF	R^2_Λ	VRF	R^2_Λ	VRF	R^2_Λ
Pre-stress							
behavior	CORT	0.433(0.1803)	0.433(0.1803)	0.372(0.1547)	0.372(0.1547)	0.402(0.1818)	0.402(0.1818)
CORT	behavior	0.060(0.0314)	0.433(0.1813)	0.039(0.020)	0.372(0.1547)	0.026(0.290)	0.402(0.1818)
behavior	heart rate	0.807(0.0928)	0.807(0.0928)	0.816(0.1116)	0.816(0.1116)	0.798(0.1793)	0.798(0.1793)
heart rate	behavior	0.119(0.0568)	0.807(0.0928)	0.069(0.0624)	0.816(0.1116)	0.071(0.0689)	0.798(0.1793)
behavior	blood pressure	0.571(0.1916)	0.571(0.1916)	0.586(0.1781)	0.586(0.1781)	0.408(0.2146)	0.408(0.2146)
blood pressure	behavior	0.081(0.0246)	0.571(0.1916)	0.073(0.0468)	0.586(0.1781)	0.011(0.0369)	0.408(0.2146)
Post-stress							
behavior	CORT	0.386(0.1889)	0.386(0.1889)	0.499(0.2095)	0.499(0.2095)	0.359(0.1190)	0.359(0.1190)
CORT	behavior	0.038(0.0248)	0.386(0.1889)	0.045(0.0984)	0.499(0.2095)	0.032(0.0273)	0.359(0.1190)
behavior	heart rate	0.913(0.0498)	0.913(0.0498)	0.984(0.0263)	0.984(0.0263)	—	—
heart rate	behavior	0.227(0.0868)	0.913(0.0498)	0.126(0.0755)	0.984(0.0263)	—	—
behavior	blood pressure	0.343(0.1041)	0.343(0.1041)	0.513(0.2050)	0.513(0.2050)	—	—
blood pressure	behavior	0.079(0.055)	0.343(0.1041)	0.160(0.1288)	0.513(0.2050)	—	—

Table 2: R^2_{indiv} *values (asymptotic standard errors) under pre-stress and post-stress conditions, for a variety of true and surrogate endpoints, using unstructured, fractional polynomial, and penalized-splines models, and based on both* VRF *and* R^2_Λ.

endpoint		unstructured		fract. pol.		pen. splines	
true	surrogate	VRF	R^2_Λ	VRF	R^2_Λ	VRF	R^2_Λ
Pre-stress							
behavior	CORT	0.433(0.1178)	0.433(0.1178)	0.372(0.1174)	0.372(0.1174)	0.402(0.1179)	0.402(0.1179)
CORT	behavior	0.060(0.0632)	0.433(0.1178)	0.039(0.0533)	0.372(0.1174)	0.026(0.0463)	0.402(0.1179)
behavior	heart rate	0.807(0.0724)	0.807(0.0724)	0.816(0.0702)	0.816(0.0702)	0.798(0.0745)	0.798(0.0745)
heart rate	behavior	0.119(0.0850)	0.807(0.0724)	0.069(0.0669)	0.816(0.0702)	0.071(0.0677)	0.798(0.0745)
behavior	blood pressure	0.571(0.1105)	0.571(0.1105)	0.586(0.1091)	0.586(0.1091)	0.408(0.1179)	0.408(0.1179)
blood pressure	behavior	0.081(0.0717)	0.571(0.1105)	0.073(0.0685)	0.586(0.1091)	0.011(0.0823)	0.408(0.1179)
Post-stress							
behavior	CORT	0.386(0.1177)	0.386(0.1177)	0.499(0.1156)	0.499(0.1156)	0.359(0.1171)	0.359(0.1171)
CORT	behavior	0.038(0.0528)	0.386(0.1177)	0.045(0.0563)	0.499(0.1156)	0.032(0.00497)	0.359(0.1171)
behavior	heart rate	0.913(0.0415)	0.913(0.0415)	0.984(0.0108)	0.984(0.0108)	—	—
heart rate	behavior	0.227(0.1063)	0.913(0.0415)	0.126(0.0868)	0.984(0.0108)	—	—
behavior	blood pressure	0.343(0.1164)	0.343(0.1164)	0.513(0.1149)	0.513(0.1149)	—	—
blood pressure	behavior	0.079(0.0709)	0.343(0.1164)	0.160(0.0947)	0.513(0.1149)	—	—