# Using earlier measures in a longitudinal sequence as a potential surrogate for a later one



DOI: 10.1016/j.csda.2009.11.024
Handle: http://hdl.handle.net/1942/10891

# Using Earlier Measures in a Longitudinal Sequence as Potential Surrogate for a Later One

**Assam Pryseley**[1] **Abel Tilahun**[1] **Ariel Alonso**[1]

**Geert Molenberghs**[1,2*]

Interuniversity Institute for Biostatistics and statistical Bioinformatics
[1] Universiteit Hasselt, Diepenbeek, Belgium
[2] Katholieke Universiteit Leuven, Leuven, Belgium

**Abstract**

The number of potential surrogate markers for clinical-trial endpoints is increasing rapidly, not in the least owing to the availability of biomarkers. At the same time, considerable development has taken place regarding statistical evaluation paradigms for such markers. As a consequence, such endpoints are given more extensive consideration for practice than previously had been the case. A particular but important instance is where the true endpoint is the ultimate assessment in a sequence of repeated measures. It is then appealing to consider earlier measures, either in isolation or several combined, as a potential surrogate endpoint. The length and cost reducing potential has to be weighed carefully against loss in precision and the risks of an inappropriate decision regarding a new compound's fate. Quantitative criteria to do so are developed, embedded in a meta-analytic framework. The methodology's behavior is assessed through simulations and applied to data from a pair of clinical trials, one in opthalmology and one in schizophrenia.

*Some Key Words:* Biomarker, Cost function, Opthalmology, Schizophrenia.

## 1 Introduction

Repeated measures of a quantitative (bio)marker are nowadays commonly obtained in clinical trials. When such measurements have the ability to predict, and/or explain a large proportion of the variability of future clinical measurement or status of a patient, then the (bio)marker may be used as a surrogate for the final measurements or status of a patient at the end of the study. If this is the case, such a (bio)marker may lead to reduction of the study's length and/or cost.

*Corresponding Author: Geert Molenberghs, I-BioStat, Center for Statistics, Universiteit Hasselt, Agoralaan, B-3590 Diepenbeek, Belgium. Email: geert.molenberghs@uhasselt.be

Surrogate-marker evaluation endeavors that have been performed thus far involved two different endpoints (Buyse *et al.* 2000, Burzykowski, Molenberghs, and Buyse 2005), where one endpoint is a candidate surrogate and the other is a true endpoint. Such endpoints may be of the same nature (e.g., both continuous, binary, or time-to-event) or of a mixed nature (e.g., an ordinal surrogate, such as tumor response, for a time-to-event endpoint, such as overall survival).

In contrast, the scenario under investigation here has only one endpoint, measured repeatedly over time. We are then interested in the predictive potential of the earlier clinical measurements for the later ones, and in particular for the last one. This can be placed within the surrogate-marker evaluation context, by considering the accumulated first few repeated measurements as potential surrogates and the outcome, for example at the final measurement occasion, as the true endpoint. Thus, for each patient, the surrogate is a vector of repeated measurements and the true endpoint is a scalar. The situation where the surrogate is a single early measurement is, of course, merely a special case.

The challenge is to determine the number of repeated measures that are required to sufficiently adequately predict the true endpoint. It is evident that collecting more repeated measurements enhances prediction. However, more repeated measurements imply longer study periods and increase cost. Thus, there must be a balance between cost and precision.

The objective of this article is threefold. First, existing surrogate-marker evaluation procedures will be tuned to accommodate the present scenario. Second, selection of an optimal number of repeated measurements will be effectuated using an objective function, designed as a weighted function of financial cost and predictive precision. The objective function allows tuning to the specific needs of a particular case study. Third, a simulation study is conducted to investigate the performance of the proposed procedure under different covariance structures for the repeated measures.

The paper is organized as follows. An introduction to the motivating studies is given in Section 2. In Section 3, we set out with a concise description of the meta-analytic approach to surrogate marker evaluation for repeated measurements using canonical correlations, as proposed by Alonso *et al* (2004), and then proceed with our modification to the scenario where early measurements on a longitudinal endpoints are treated as a surrogate for the final measurement. Section 4 provides,

from a theoretical point of view, the performance of an objective function for two important special cases. Section 5 provides details on the design and results of our simulation study, and provide a perspective on the conclusions that can be drawn from it. In Section 6, we briefly introduce a constrained maximization problem. Section 7 contains the results of the case studies' analysis.

# 2 Motivating Case Studies

## 2.1 Age-related Macular Degeneration Study

This is a clinical trial involving patients with age-related macular degeneration (ARMD), a condition in which patients progressively lose vision. Overall, 1186 patients from 114 sites participated in the trial. Patients' visual acuity was assessed using standardized vision charts displaying lines of five letters of decreasing size that patients had to read from top to bottom. Visual acuity is captured as the number of letters correctly read. The binary indicator for treatment is set to $Z = -1$ for placebo and $Z = 1$ for treatment. In the analysis, the sites at which patients were treated will be considered as units of analysis. Some of the sites participating in the trial enrolled patients only to one of the two treatment arms and were excluded from further considerations. A total of 82 sites were thus available for analysis, with the number of individual patients per center ranging from 2 to 19, totaling to 424 patients overall.

## 2.2 A Meta-analysis of Five Clinical Trials in Schizophernia

The data come from a meta-analysis of five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. The treatment indicator for risperidone versus conventional treatment will be denoted by $Z$. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both 'negative' and 'positive' symptoms. Negative symptoms are characterized by deficits in cognitive, affective, and social functions; for example, poverty of speech, apathy, and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations, and disorganized thinking, which are superimposed on mental status (Kay, Fiszbein, and Opler 1987). Several measures can be considered to asses a patient's global condition. One useful and sufficiently sensitive assessment scale is the Positive and Negative Syndrome Scale (PANSS) (Kay, Opler, and

Lindenmayer 1988). The PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay, Opler, and Lindenmayer 1988). We will apply the methods proposed to the repeatedly measured PANSS outcome. The data are made up of five trials, all containing information on the treating investigators, necessary for defining the units for analysis.

# 3 Longitudinal Endpoints and Surrogacy

We set out, in Section 3.1 with a brief description of the existing meta-analytic approach to surrogate marker evaluation for repeated measurements, using canonical correlations as introduced by Alonso *et al.* (2004). This will be followed in Section 3.2 by a version tailored to the needs of our goal, i.e., the determination of an optimal number of repeated measurements required to accurately predict the true endpoint. In Section 3.3, we zoom in on the development of an objective function.

## 3.1 Canonical Correlation Approach for Two Repeatedly Measured Endpoints

We shall assume that information from $i = 1, \ldots, N$ trials is available, in the $i^{th}$ of which, $j = 1, \ldots, n_i$ subjects are included. We shall further denote the time points at which each subject in trial $i$ is measured as $t_{ik}$. If $T_{ijk}$ and $S_{ijk}$ denote the associated true and surrogate endpoints at time $k$, respectively, and $Z_{ij}$ is a binary indicator variable for treatment, Alonso *et al.* (2004) proposed the following joint model for both responses:

$$S_{ijk} = \mu_{S_i} + \alpha_i Z_{ij} + \theta_{S_i} t_{ik} + \varepsilon_{S_{ijk}}, \quad (1)$$

$$T_{ijk} = \mu_{T_i} + \beta_i Z_{ij} + \theta_{T_i} t_{ik} + \varepsilon_{T_{ijk}}, \quad (2)$$

where $\mu_{S_i}$ and $\mu_{T_i}$ are trial-specific intercepts, $\alpha_i$ and $\beta_i$ are trial-specific effects of treatment $Z_{ij}$ on the two endpoints, and $\theta_{S_i}$ and $\theta_{T_i}$ are fixed trial-specific time effects. The vectors $\varepsilon_{S_{ij}}$ and $\varepsilon_{T_{ij}}$ are assumed to have a multivariate normal distribution with mean zero and variance-covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{SS_i} & \Sigma_{ST_i} \\ \Sigma_{TS_i} & \Sigma_{TT_i} \end{pmatrix}. \quad (3)$$

In (3), $\Sigma_{TT_i}$ and $\Sigma_{SS_i}$ denote the variance-covariance matrices associated with the true and the surrogate endpoints, respectively, and $\Sigma_{TS_i} = \Sigma_{ST_i}^T$ contains the covariances between the measure-

ments for the true and the surrogate endpoints. In some practical settings, $\Sigma_i$ can be modeled as the Kronecker product of a general correlation matrix that captures the association within the sequences and an unstructured $2 \times 2$ matrix that captures the association between the sequences (Galecki 1994).

One might be interested in studying how an individual's surrogate score is predictive of the true score, which is referred to as individual-level surrogacy. Due to the longitudinal nature of the endpoints, Alonso *et al.* (2004) extended the ideas of Buyse *et al.* (2000) for capturing individual-level surrogacy, based on coefficients of determination, to a multivariate version using the concept of canonical correlation. Based on model (3), these authors obtained the canonical correlations, $\rho_{ik}$ from $\Sigma_{TT_i}^{-1}\ \Sigma_{TS_i}\Sigma_{SS_i}^{-1}\Sigma_{TS_i}^{T}$ and proposed a family of measures to evaluate surrogacy at the individual level. This so-called $\Omega$ family is defined as

$$\Omega = \left\{ \vartheta : \vartheta = \sum_i \sum_k \alpha_{ik}\rho_{ik}^2, \quad \text{where:} \quad \alpha_{ik} > 0 \quad \forall (i,k), \quad \sum_i \sum_k \alpha_{ik} = 1 \right\}. \tag{4}$$

An important member of the $\Omega$ family is the *Variance Reduction Factor* (VRF) originally introduced by Alonso *et al.* (2003) and defined as

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TT_i}) - \text{tr}(\Sigma_{T|S_i})\}}{\sum_i \text{tr}(\Sigma_{TT_i})}, \tag{5}$$

where $\Sigma_{T|S_i}$ denotes the conditional variance-covariance matrix of $\varepsilon_{T_{ij}}$ given $\varepsilon_{S_{ij}}$. Intuitively, (5) quantifies how much of the total variability in the true endpoint is explained by adjusting for the treatment effects and the (repeated measurements on) the surrogate endpoints. Values close to 1 indicate that the surrogate is a 'good' predictor for the true endpoint while values close to 0 indicate a 'poor' predictor. Evidently, values for $VRF_{\text{ind}}$ have to be complemented with biopharmaceutical, regulatory, and other expert opinion.

### 3.2 Optimal Number of Repeated Measurements

We are interested in predicting a patient's outcome at a specified point in time from an accumulated number of repeated measurements at earlier times. To this end, let us denote by $Y_{ijk}$ the $k^{th}$ measurement on subject $j$ in trial $i$. We shall further assume that the following model holds:

$$Y_{ijk} \quad = \quad (\beta_0 + b_{1i}) + (\beta_1 + b_{2i})Z_{ij} + \beta_2 t_{ik} + \beta_3 Z_{ij} t_{ik} + \varepsilon_{ijk}, \tag{6}$$

where $Z_{ij}$ and $t_{ik}$ are defined as before, $(b_{1i}, b_{2i})$ are trial-specific (random) effects, assumed to follow a zero-mean normal distribution with covariance matrix $D_L$, and the error vector $\varepsilon_{ij}$ is assumed zero-mean normally distributed with covariance matrix $\Sigma_L$.

Note that there are some important differences between models (6) and (1)–(2). Indeed, (1)–(2) is based on two different, repeatedly measured endpoints, while (6) is based on a single sequence of which earlier components act as surrogates for later ones. In many applications, assuming a constant treatment effect over time will be unrealistic. We assume a linear treatment effect over time, constant across trial, but extension of which is straightforward (Alonso *et al.* 2004).

Let us formally define our surrogate and true endpoints, based on (6). Supposed we intend to investigate whether the first $m$ accumulated measurements, where $1 \leq m \leq K-1$, constitute a good set of predictors for the outcome measured at time $K$. Hence, $S_{ijk} = Y_{ijk}$, $k = 1, \ldots, m$, and $T_{ij} = Y_{ijK}$, leading to the following possible model:

$$\begin{aligned} S_{ijk} &= \mu^*_{S_i} + \alpha^*_i Z_{ij} + \beta_2 t_{ij} + \beta_3 t_{ij} Z_{ij} + \varepsilon_{S_{ijk}}, \\ T_{ij} &= \mu^*_{T_i} + \beta^*_i Z_{ij} + \varepsilon_{T_{ij}}, \end{aligned} \tag{7}$$

where the random-effects vector $(\mu^*_{S_i}, \mu^*_{T_i}, \alpha^*_i, \beta^*_i)$ is assumed to be zero-mean normally distributed with covariance matrix:

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{8}$$

and the $m+1$ dimensional error vector $(\varepsilon^T_{S_{ij}}, \varepsilon_{T_{ij}})^T$ is zero-mean normally distributed with variance-covariance matrix $\Sigma$. The analyst will have to decide on an appropriate structure for $\Sigma$, which can be usefully partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{ST} \\ \Sigma_{TS} & \Sigma_{TT} \end{pmatrix}, \tag{9}$$

where $\Sigma_{ST}$ is a vector and $\Sigma_{TT} = \sigma_{TT}$ is a scalar. Similar to Buyse *et al.* (2000), linear mixed-effects methodology can be used for parameter estimation and inferences (Verbeke and Molenberghs 2000).

Applying the $VRF_{\text{ind}}$ in this setting leads to

$$VRF_{\text{ind}} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{T|S})}{\text{tr}(\sigma_{TT})}, \tag{10}$$

where $\sigma_{T|S}$ denotes the conditional variance of $T_{ij}$ given the surrogates: $\sigma_{T|S} = \sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. We can re-express (10) as

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

Note that $VRF_{\text{ind}} = 0$ if and only if $\Sigma_{ST} = 0$, i.e., if and only if when $T_{ij}$ and $\widetilde{S}_{ij}$ are independent, where $\widetilde{S}_{ij}$ is the vector grouping the surrogate measures.

The bivariate linear mixed models (BLMM) used by Buyse *et al.* (2000) are very flexible models, however, obtaining convergence with such models is a non-trivial task. To address this, Tibaldi *et al.* (2003) and Tilahun *et al.* (2007) proposed and studied a number of simplified fitting strategies. Here, we will proceed by so-called bivariate general linear models (BGLM), essentially a two-stage, fixed-effects version of the BLMM, which has shown good performance in both statistical and computational terms. Precisely, in the first stage of a BGLM, the effects $\mu_{S_i}^*$, $\mu_{T_i}^*$, $\alpha_i^*$, and $\beta_i^*$ in (7) are considered fixed. The error term $(\varepsilon_{S_{ij}}^T, \varepsilon_{T_{ij}})^T$ is assumed to follow a zero-mean normal distribution with covariance matrix $\Sigma$.

### 3.3 Cost Function and Optimal Number of Measurements

To determine the optimal number of measurements ($m_o$), we will consider the following cost function, introduced by Winkens *et al.* (2005):

$$FC = NC_1 + NKC_2. \tag{12}$$

Here, $FC$ represents the fixed total financial cost, $N$ is the total number of patients in the study, $K$ is the number of planned repeated measurements per subjects, $C_1$ is the cost of recruiting a patient to the study, and $C_2$ is the cost per measurement and per subject. Let $R = C_1/C_2$ be the ratio of both costs; usually the cost of recruiting a patient to the study is higher than the cost per measurement, i.e., $R > 1$. We can then re-write (12) as $FC = NC_2(R + K)$. Suppose now that, instead of taking $K$ measurements, we take $m$, $1 \leq m \leq K - 1$, measurements and use this information to predict the outcome at the $K^{th}$ time point, the financial cost for the $m$ measurements is then given by $FC(m) = NC_1 + NmC_2$. Thus, the proportion of the total financial cost required to take $m$ measurement is $PFC(m) = (R + m)/(R + K)$. It is easy to show that the variance of the prediction, based on $m$ observations, of the outcome at the last time point takes

the form $[1 - VRF_{\text{ind}}(m)]\sigma_{TT}$. Note further that $\sigma_{TT}$ is constant, irrespective of the number of repeated measurements used as a surrogate; thus a standardized version of the prediction variance, $1 - VRF_{\text{ind}}(m)$, will be used. Finally, a weighted linear combination of the prediction variance and the financial cost can be used to define an objective function as shown in (13), with weights $w_1$ and $(1 - w_1)$, respectively. An advantage of standardizing the prediction variance and financial cost for a given number of repeated measurements $m$ is the relative ease of specifying $w_1$, compared to using the non-standardized versions:

$$CPR_0(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + (1 - w_1) \cdot \frac{R + m}{R + K}, \tag{13}$$

The quantity $CPR_0(m)$ balances the lack of surrogacy, $1 - VRF_{\text{ind}}(m)$, on the one hand, and the proportion of total financial cost required to take $m$ measurements, $(R+m)/(R+K)$, on the other hand. Retaining more measurements reduces the first term, because the VRF will go up, but at the same time leads to an increase in the cost term. The relative importance attributed to the terms is captured by the weight $w_1$, with a user-assigned value between 0 and 1. The number $m_o$ is determined as that minimizing $CPR(m)$.

Let us consider some extensions. The objective function assumes that the cost of each measurement is the same, which may be unrealistic for some situations; for example, when patients have to stay in a hospital or health institute, where the waiting time may incur additional costs, a feature not accommodated by (13). One can therefore elect to introduce a third term accounting for time lag:

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R + m}{R + K} + w_3 \cdot \frac{t_m - t_0}{t_k - t_0}, \tag{14}$$

If the repeated measures are equidistant with time lag $\triangle$, then $t_m = t_0 + \triangle M$ and $t_k = t_0 + \triangle K$. Hence, (14) takes the form

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R + m}{R + K} + w_3 \cdot \frac{M}{K}. \tag{15}$$

If in addition we assume that the waiting cost for the first measurement is zero, then:

$$CPR_{II}(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R + m}{R + K} + w_3 \cdot \frac{M - 1}{K}. \tag{16}$$

These objective functions assume that the cost is constant across treatment arms, whether of a placebo, standard-therapy, or experimental nature. When deemed unrealistic, appropriate modifications can be implemented. Arguably, the choice of a cost function will have to balance simplicity

with it being a realistic representation of reality. In what follows, objective function (13) will be employed, unless otherwise stated.

Starting from 13, it is insightful to derive what is needed to have the VRF go up by a given amount $\Delta$. If the CPR is to remain identical, them the new optimum number of measurements $m_1$, is an increase over the original optimum $m_0$, as follows:

$$m_1 = m_0 + \frac{w_1}{1 - w_1}(R + K)\Delta.$$

Obviously, the smaller the weight, the less sensitive the number of measurements becomes. It is informative to derive the corresponding cost increase, starting from 12:

$$N\frac{w_1}{1 - w_1}(R + K)\Delta C_2.$$

Also here, the dampening effect of a smaller $w_i$ is clearly visible.

# 4 Some Important Special Cases

In this section, we aim to aid understanding of the nature of the cost functions through theoretical considerations for two special, important cases.

## 4.1 Compound Symmetry Structure

Assume that the covariance structure of (6) is compound symmetry, i.e., $\Sigma_L = \sigma(1-\rho)I_K + \sigma\rho J_K$, where $\sigma$ denotes the variance of the response at each time point, $\rho$ is the correlation between two observations, $I_K$ is a $K$-dimensional identity matrix and $J_K$ is a $K$-dimensional square matrix of ones. It is easy to show that, in this setting,

$$VRF_{\text{ind}}(m) = \frac{m\rho^2}{1 + (m-1)\rho}.$$

Let us study the predictive characteristics of this case. It follows that $VRF_{\text{ind}}(m)$ is an increasing function of $m$ as far as $\rho \neq 0, 1$ and, therefore, the more observations we include in $\widetilde{S}_{ij}$, the more precise our prediction of $T_{ij}$ will be. Turning to $\rho$, the question is how the correlation influences the amount of information that $\widetilde{S}_{ij}$ brings about $T_{ij}$. To usefully study this, let us calculate the

additional information that one extra observation will bring, quantified using the ratio:

$$g(\rho) = \frac{VRF_{\text{ind}}(m+1)}{VRF_{\text{ind}}(m)} = \left(\frac{m+1}{m}\right)\left(\frac{1+(m-1)\rho}{1+m\rho}\right).$$

Some elementary calculations show that $g(\rho)$ is a decreasing function of $\rho$ and therefore, the higher the correlation the less we gain by taking additional observations, rather an intuitive result. Indeed, if the correlation is very high, then all the measurements are nearly deterministically related, and having observed one or a few of them will allow us to predict with high precision all the others. For instance, in the extreme case when $\rho = 1$ the $VRF_{\text{ind}}(m+1) = VRF_{\text{ind}}(m)$ for all $m$ and the first observation will be sufficient to predict the true endpoint without error.

Coherent with the nature of compound symmetry, the position in the sequence of the $m$ observations that constitute the surrogate is totally irrelevant. It is easy to show that in this setting the $CPR$ function takes the form

$$CPR(m) = w_1 \cdot \frac{(1-\rho)(1+m\rho)}{1+(m-1)\rho} + (1-w_1) \cdot \frac{R+m}{R+K}, \tag{17}$$

of which the extremes are easy to determine: (17) reaches its minimum at $m_+$ and $m_-$ when $\rho > 0$ and $\rho < 0$ respectively, where

$$m_\pm = -\left(\frac{1-\rho}{\rho}\right) \pm \sqrt{\frac{w_1(R+K)(1-\rho)}{1-w_1}}. \tag{18}$$

Obviously, in many practical situations, $m_\pm$ will not be integers, in which case they will have to be rounded. There is also a possibility for $m_\pm$ to assume a negative value for some combinations of $K$, $\rho$, $R$, and $w_1$. When this happens, $m_\pm$ should be set to one.

Zooming in on $m_+$ reveals that, when less weight is assigned to the precision part of the cost function, an increase in $R$ has little influence on $m_+$ but its influence increases as more weight is assigned to precision. This is to be expected because when the cost of recruiting patients is much higher than taking more measurements on subjects, the obvious way to increase precision is through taking more measurement per subject. An increase in the correlation $\rho$ between measurement leads to a decrease in $m_+$ when the weight assigned to precision is small to moderate. When the weight increases, the value of $m_+$ increase for $\rho$ in $[0; 0.5]$ and decreases in $[0.5; 1]$. Also, a increase in $K$ generally leads to a slight increase in $m_+$.

### 4.2 First-order Auto-regressive Process

Another association structure frequently encountered in longitudinal data is the first-order auto-regressive one, with $\rho^t$ the correlation between two measurements, $t$ time units apart. In this case, $\Sigma_{SS}$ is also an $(m \times m)$ AR(1) matrix, $\Sigma_{ST} = \Sigma_{TS}^T = \rho^{K-m}\delta_1^T$ with $\delta_1^T = (\rho^{m-1}, \ldots, 1)$ and $\sigma_{TT} = \sigma$. It then follows that $VRF_{\text{ind}}(m) = \rho^{2(K-m)}\sigma\delta_1^T\Sigma_{SS}^{-1}\delta_1$. Further, using the expression for the inverse of an AR(1) matrix (Graybill 1983), one can prove that $\sigma\delta_1^T\Sigma_{SS}^{-1}\delta_1 = 1$ and therefore $VRF_{\text{ind}}(m) = \rho^{2(K-m)}$. Like in the compound-symmetry case, here the $VRF_{\text{ind}}(m)$ is an increasing function of $m$. However, unlike before, it is also an increasing function of $\rho$, implying that the higher $\rho$, the more advantageous it is to include more observations into the surrogate. This is again a very intuitive result. This is intuitively plausible because, under AR(1), the correlation decreases rapidly with time lag; hence it is recommendable to consider surrogate outcomes that are collected sufficiently closely to the true endpoint. More generally, the position of the surrogate measures within the sequence of repeated measures is now relevant. For instance, if we now consider as the surrogate marker a sub-sequence of $m$ observations starting at time point $s+1$, then $VRF_{\text{ind(s+1)}}(m) = \rho^{2(K-s-m)}$. Obviously, $VRF_{\text{ind(s+1)}}(m) \geq VRF_{\text{ind}}(m)$, for $s \geq 1$, and therefore considering $m$ observations closer to the true endpoint will result in a surrogate with more predictive power. In this scenario, the $CPR$ function takes the form:

$$CPR(m) = w_1 \cdot \left(1 - \rho^{2(K-m)}\right) + (1 - w_1) \cdot \frac{R+m}{R+K}. \tag{19}$$

Interestingly, (19) does not reach its minimum value in the interval $(1, K-1)$ and therefore $CPR(m)$ will always lead to choosing the first observation only if the cost is the impelling criterion or choosing the entire $K-1$ sequence if prediction is the more important factor. This result also holds if the longitudinal surrogate sequence is started at a time point different from the first one. Thus, the $CPR(m)$ seems to indicate that in this scenario the surrogate should contain one observation only and therefore, the most rational choice would be to consider a value sufficiently close to the true endpoint so that a reasonable level of precision can be achieved in the prediction. Obviously, the closer this observation is to the true endpoint the better the prediction will be but the longer we will have to wait. A compromise between these two considerations should be found in this setting using external elements such as, for example, expert opinion.

# 5 Simulation Study

Even though the previous results are enlightening, not all cases are analytically tractable. Moreover, even in those cases where analytic results are obtainable it is still of great interest to study the performance of the proposed method when parameters have to be estimated. A simulation study was performed to investigate further these issues, with focus on the two association structures of Section 4.

## 5.1 Data Generation

Equally spaced longitudinal data were generated based on (6) and using a two-stage approach. In the first stage, random trial-specific intercepts and treatment effects, $b_{1i}$ and $b_{2i}$ respectively, were generated from a zero-mean normal distribution with covariance matrix

$$D_L = \begin{pmatrix} 1.5 & 2.098 \\ 2.098 & 3.26 \end{pmatrix}.$$

Additionally, error terms $\varepsilon_{ijk}$ were generated from a zero-mean normal distribution with covariance matrix $\Sigma_L$, either first-order autoregressive, AR(1), or compound symmetry, CS. The variance in $\Sigma_L$ was assumed constant and the correlation between successive measurements was set to either 0.3, 0.6, or 0.9. The fixed-effects vector was set to $\boldsymbol{\beta}^T = (2.5, 4.3, 0.78, 3.5)$. Using these, the outcomes were obtained from (6).

The data generation scheme discussed earlier assumes that the treatment-by-time interaction is constant across trials. To increase flexibility, a more general framework, where the treatment effect is allowed to randomly vary over time and across trials was adopted. The first stage now involved generation of random trial-specific time effects and random slopes, in addition to random trial-specific intercepts and treatment effects, $b_{1i}$ and $b_{2i}$, from a zero-mean normal distribution with covariance matrix

$$D_L = \begin{pmatrix} 1.0 & 0.8 & 0.00 & 0.00 \\ 0.8 & 1.0 & 0.00 & 0.00 \\ 0.0 & 0.0 & 1.00 & 0.95 \\ 0.0 & 0.0 & 0.95 & 1.00 \end{pmatrix}.$$

The error terms were, again, generated from a zero-mean normal distribution with AR(1) or CS

covariance matrix $\Sigma_L$, The outcome vector $Y_{ijk}$ then takes the form:

$$Y_{ijk} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})Z_{ij} + (\beta_2 + b_{2i})t_{ij} + (\beta_3 + b_{3i})Z_{ij}t_{ij} + \varepsilon_{ijk}.$$

The number of trials was set to either 10, 20, 30, or 40. Two sets of trial sizes were considered. The first set of smaller trial sizes consists of 20, 40, and 60 subjects per trial. The second set of larger trial sizes consists of 100, 200, and 300 subjects per trial. The simulation consists of a full combination of the specified correlation values, covariance matrix structures, number of trials, and trial sizes. For each combination, 100 datasets (samples) where generated as described in Section 5.1, analyzed and the optimal number of measurements determined as described in Section 3.

In principle, simulations based on 100 runs are in jeopardy of large Monte Carlo errors. However, because we predominantly determine the optimal number of measurements, a discrete quantity, there is little gain to be expected from increasing the number of runs.

### 5.2 Simulation Study Results

The results of the simulation for the case of $R = 4$ and $K = 10$ are summarized in Tables 1–4. In the tables, $VRF_{\text{ind}}(m_o)$ is the usual individual-level surrogacy for the optimal number of measurements, while $VRF_{\text{ind}}(K-1)$ corresponds to the entire $K-1$ sequence being used as a surrogate. Furthermore, $f$ represents the percentage of datasets that resulted in a given $m_o$ as the optimal number of measurements. The weight, $w_1$, was set to either 0.3, 0.5, or 0.7.

Let us focus on the first data-generation scheme, where the treatment-by-time interaction is assumed constant across trials. We learn that the $VRF_{\text{ind}}(m)$ increases with increasing number of repeated measurements. When the data are generated under AR(1) but analyzed using an unstructured covariance matrix, the optimal number of time points was chosen to be either 1 or 9, depending on the weights assigned. When the correlation was set to 0.9, assigning more weight to precision or equal weights to both precision and financial cost requires all 9 repeated measurements to minimize the objective function. For the other possible values of the correlation, i.e., 0.30, 0.60, or 0.71, if more weight is assigned to financial cost or equal weights are assigned to financial cost and precision then the optimum simply is the first measurement only. However, the entire sequence is needed when progressively more weight is assigned to the precision. This result is in agreement with

Section 4, where we have shown that, under AR(1), $CPR(m)$ does not reach its minimum value in the interval $(1, K-1)$ and therefore it will always lead to taking either only one observation or the entire $K-1$ subsequence. Hence, this result carries over to the simulation setting, in spite of the added variability coming from parameter estimation.

When the data are generated using CS and analyzed with either unstructured or CS (Table 2), then 1, 2, 3, or 4 repeated measurements may be required to predict the outcome at the last time point, with differing percentages of the sample depending on the weight assigned. When less weight is assigned to precision, the first observation is selected and the optimal number of measurements equals one, for both CS and unstructured.

Note that, in Table 2, missing entries are not due to convergence issues, for example. Actually those spaces are left for conveniently putting the results for CS and UN in one table. For example, for CS structure with a weight of 0.7 and correlation of 0.71, when the analysis was done with a CS, time point 3 was selected as optimal with 100 percent of the samples. Whereas, when the analysis is conducted with UN, time points 2 and 4 were also selected as optimal with percentages of 14 and 6 respectively. These two time points were not picked before and hence the space corresponding to time points 2 and 4 is left blank in the columns corresponding to CS. We have added this explanation to the manuscript.

In the second data-generation scheme, where treatment effects are allowed to vary, the same results followed, for both AR(1) and CS.

We also gave some consideration to the Toeplitz, or banded, structure, where the correlation between pairs of measurements varies with the time lag between them, in an unstructured way, but is independent of the actual times at which the measurements are taken. Furthermore, an AR(1)-type structure was assumed where the decline in autocorrelation is expressed in terms of the square root of the time lag, denoted by AR(1)-Sq. The results are summarized in Tables 3–4. For the Toeplitz structure up to five time points and for the unstructured matrix up to six time points were selected as optimum, depending on the weight assigned to the precision part of the cost function. For the AR(1)-Sq structure, the optimal time point swings between taking the first measurement or the entire sequence. However, it picks the first time point as optimal more often, except when

Table 1: *Simulation study. Results for the optimal number of measurements with AR(1). ($\rho$: correlation between successive time measurements; $w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; $f$: percentage of datasets resulting in $m_o$ is 100% in all cases.)*

| | | $VRF_{\text{ind}}(m)$ | | | | $VRF_{\text{ind}}(m)$ | |
|---|---|---|---|---|---|---|---|
| $w_1$ | $m_o$ | as AR(1) | as CS | $w_1$ | $m_o$ | as AR(1) | as CS |
| $\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.09$ | | | | $\rho = 0.71$ & $VRF_{\text{ind}}(K-1) = 0.50$ | | | |
| 0.7 | 1 | 0.00003 | 0.0006 | 0.7 | 9 | 0.50 | 0.50 |
| 0.5 | 1 | 0.00003 | 0.0006 | 0.5 | 1 | 0.0032 | 0.0032 |
| 0.3 | 1 | 0.00003 | 0.0006 | 0.3 | 1 | 0.0032 | 0.0032 |
| $\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.36$ | | | | $\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.81$ | | | |
| 0.7 | 9 | 0.36 | 0.42 | 0.7 | 9 | 0.81 | 0.81 |
| 0.5 | 1 | 0.07 | 0.07 | 0.5 | 9 | 0.81 | 0.81 |
| 0.3 | 1 | 0.07 | 0.07 | 0.3 | 1 | 0.15 | 0.15 |

the weight assigned to precision is as high as 70% and correlation values are 0.60 and 0.90. For a correlation of 0.30, it invariably picks the first time point only, even when the weight is as high as 70%.

## 6 Constrained Maximization

There are circumstances in which clinical trials are faced with budget constraints and yet are expected to produce acceptable results. This predicament motivates the use of constraint maximization to arrive at an optimal number of subjects and/or repeated measures per subject, thereby not exceeding the budget available. Translated to our setting, we aim at maximizing the individual level surrogacy measure, subject to cost and time constraints. We first maximize $VRF_{\text{ind}}(m)$ subject to $(R+m)/(R+K) \leq \delta_1$ and then later subject to two constraints: $(R+m)/(R+K) \leq \delta_1$ and $(t_m - t_0)/(t_k - t_0) \leq \delta_2$, where both $\delta_1$ and $\delta_2$ assume values between zero and one.

Without loss of generality, if we assume that the measurements are equally spaced with fixed time interval $\triangle$, then $t_m = t_0 + \triangle M$ and $t_k = t_0 + \triangle K$ and hence the second constraint reduces to $M/K \leq \delta_2$. Using a Lagrange multiplier for the first optimization problem, one can show that, for CS with positive $\rho$, the optimal number of repeated measures required for a percentage budget of

Table 2: *Simulation study. Results for the optimal number of measurements with CS. ($\rho$: correlation between successive time measurements; $w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; $f$: percentage of datasets resulting in $m_o$.)*

| | | as CS | | as UN | |
|---|---|---|---|---|---|
| $w_1$ | $m_o$ | $VRF_{\text{ind}}(m)$ | $f$ | $VRF_{\text{ind}}(m)$ | $f$ |
| | | $\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.24$ | | | |
| 0.7 | 1 | 0.11 | 18 | 0.10 | 18 |
| 0.7 | 2 | 0.14 | 6 | 0.12 | 34 |
| 0.7 | 3 | 0.16 | 60 | 0.17 | 22 |
| 0.7 | 4 | 0.19 | 16 | 0.19 | 26 |
| 0.5 | 1 | 0.09 | 100 | 0.09 | 100 |
| 0.3 | 1 | 0.09 | 100 | 0.09 | 100 |
| | | $\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.56$ | | | |
| 0.7 | 3 | 0.49 | 60 | 0.48 | 62 |
| 0.7 | 4 | 0.52 | 40 | 0.51 | 38 |
| 0.5 | 1 | 0.37 | 30 | 0.37 | 18 |
| 0.5 | 2 | 0.44 | 70 | 0.43 | 82 |
| 0.3 | 1 | 0.36 | 100 | 0.36 | 100 |
| | | $\rho = 0.71$ & $VRF_{\text{ind}}(K-1) = 0.68$ | | | |
| 0.7 | 2 | | | 0.58 | 14 |
| 0.7 | 3 | 0.62 | 100 | 0.62 | 80 |
| 0.7 | 4 | | | 0.64 | 6 |
| 0.5 | 3 | | | 0.62 | 70 |
| 0.5 | 4 | | | 0.64 | 6 |
| 0.5 | 1 | 0.51 | 30 | | |
| 0.5 | 2 | 0.58 | 70 | 0.57 | 24 |
| 0.3 | 1 | 0.50 | 100 | 0.50 | 100 |
| | | $\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.89$ | | | |
| 0.7 | 2 | 0.85 | 100 | 0.85 | 100 |
| 0.5 | 1 | 0.81 | 100 | 0.81 | 100 |
| 0.3 | 1 | 0.81 | 100 | 0.81 | 100 |

$\delta_1$ is given as:

$$M = \begin{cases} \delta_1(R+K) - R & \text{if } (R+1) - \delta_1(R+k) \leq \frac{1}{\rho}, \\ 2\left(\frac{1-\rho}{\rho}\right) - \delta_1(R+K) + R & \text{if } (R+1) - \delta_1(R+k) \geq \frac{1}{\rho}. \end{cases}$$

Table 3: *Simulation study. Results for the optimal number of measurements with: unstructured covariance and Toeplitz correlation structure with slowly declining correlation ( $w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; $f$: percentage of datasets resulting in $m_o$.)*

| $w_1$ | $m_o$ | $VRF_{\text{ind}}(K-1)$ | $f$ |
|---|---|---|---|
| Unstructured | | | |
| $VRF_{\text{ind}}(K-1) = 0.995$ | | | |
| 0.1 | 1 | 0.53 | 100 |
| 0.3 | 1 | 0.53 | 100 |
| 0.5 | 4 | 0.86 | 92 |
| 0.5 | 5 | 0.91 | 8 |
| 0.7 | 6 | 0.96 | 100 |
| 0.6 | 4 | 0.86 | 29 |
| 0.6 | 5 | 0.91 | 57 |
| 0.6 | 6 | 0.96 | 14 |
| Toeplitz | | | |
| $VRF_{\text{ind}}(K-1) = 0.75$ | | | |
| 0.1 | 1 | 0.15 | 100 |
| 0.3 | 2 | 0.16 | 80 |
| 0.3 | 3 | 0.22 | 20 |
| 0.5 | 4 | 0.38 | 100 |
| 0.6 | 4 | 0.38 | 98 |
| 0.6 | 5 | 0.42 | 2 |
| 0.7 | 5 | 0.42 | 100 |

In a similar manner, for AR(1) with $\rho \geq 0$, the optimal number of repeated measures for a given percentage of the budget is $M = \delta_1(R+K) - R$. If we now maximize the association measure subject to both budget and time constraint, we find $M = \min[\delta_1(R+K) - R, \delta_2 K]$ for the optimal number of repeated measures for both CS and AR(1).

To enhance insight, we carried out a limited set of simulations for both AR(1) and CS. The simulation has revealed that as $R$ increases, the optimal $M$ diminishes. Results are summarized in Table 5. This is in line with intuition because the total cost and the number of patients in the study

Table 4: *Simulation study. Results for the optimal number of measurements with: AR(1) with square root of time lag analyzed as conventional AR(1). ( $w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; $f$: percentage of datasets resulting in $m_o$.)*

| $w_1$ | $m_o$ | $VRF_{\text{ind}}(K-1)$ | $f$ |
|---|---|---|---|
| AR(1)-Sq | | | |
| $\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.22$ | | | |
| 0.1 | 1 | 0.0016 | 100 |
| 0.3 | 1 | 0.0016 | 100 |
| 0.5 | 1 | 0.0016 | 100 |
| 0.6 | 1 | 0.0016 | 100 |
| 0.7 | 1 | 0.0016 | 100 |
| AR(1)-Sq | | | |
| $\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.50$ | | | |
| 0.1 | 1 | 0.052 | 100 |
| 0.3 | 1 | 0.052 | 100 |
| 0.5 | 1 | 0.052 | 100 |
| 0.6 | 9 | 0.052 | 100 |
| 0.7 | 9 | 0.052 | 100 |
| AR(1)-Sq | | | |
| $\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.86$ | | | |
| 0.1 | 1 | 0.21 | 100 |
| 0.3 | 1 | 0.21 | 100 |
| 0.5 | 1 | 0.21 | 100 |
| 0.6 | 1 | 0.21 | 100 |
| 0.7 | 9 | 0.86 | 100 |

are fixed and hence to maintain a low cost, the only option is to reduce the number of repeated measures. It also follows that, for some values of $R$, it is not possible to obtain a value of $M$ for which the percentage of cost incurred is lower than the specified $\delta$ value. In such cases, only the first time point or the entire sequence could be taken, depending on the magnitude of $M$. In this context, it is also worth noting that, although there is no difference in the optimal number of repeated measures for CS and AR(1), the same number of repeated measures in the two covariance structures will nevertheless not yield identical $VRF_{\text{ind}}(m)$ values.

Table 5: *Simulation study for constraint maximization. Results for the optimal number of measurements for $\rho$= 0.3 with CS and AR(1). ($\delta$: percentage of cost available; R: cost ratio; $m_o$: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements $m_o$.)*

| CS | | | | AR(1) | | | |
|---|---|---|---|---|---|---|---|
| $\delta$ | $R$ | $m_0$ | $VRF_{\text{ind}}(m)$ | $\delta$ | $R$ | $m_0$ | $VRF_{\text{ind}}(m)$ |
| 0.2 | 1 | 1 | 0.10979 | 0.2 | 1 | 1 | 2.29E-14 |
| 0.3 | 1 | 2 | 0.13882 | 0.3 | 1 | 2 | 4.44E-09 |
| 0.4 | 1 | 3 | 0.16273 | 0.4 | 1 | 3 | 4.11E-08 |
| 0.5 | 4 | 3 | 0.17758 | 0.5 | 4 | 3 | 4.11E-08 |
| 0.6 | 4 | 4 | 0.19843 | 0.6 | 4 | 4 | 2.85E-07 |
| 0.8 | 4 | 7 | 0.21728 | 0.8 | 4 | 7 | 0.000584203 |
| 0.6 | 10 | 2 | 0.19843 | 0.6 | 10 | 2 | 4.44E-09 |
| 0.7 | 10 | 4 | 0.20691 | 0.7 | 10 | 4 | 2.85E-07 |
| 0.8 | 10 | 6 | 0.21728 | 0.8 | 10 | 6 | 4.68688E-05 |
| 0.9 | 10 | 8 | 0.22203 | 0.9 | 10 | 8 | 0.007548459 |

# 7 Application to the Case Study

The two case studies introduced in Section 2 are analyzed here and the results displayed in Tables 6 and 7, respectively. For the data coming from the opthalmology experiment, measurements of visual acuity were taken at baseline and every sixth week there after up to $54^{th}$ week giving 10 repeated measures. For the schizophrenia study, the PANSS values were measured at five different time points, taken at the baseline and every two weeks thereafter. In both cases, the objective is to predict the ultimate measurement using earlier ones from the sequence, thereby accounting for cost. In both cases, an unstructured variance-covariance matrix fits the data best.

Now focusing attention on the data coming from the opthalmology experiment, we find that, with increasing weight attributed to precision: the first one; the first and the second; the first, the second, and the third; the first eight; or all nine time points were required to optimally predict the final measurement. Note that one time unit corresponds to 6 weeks. Thus, for example, taking the first three time points amounts to using measurements from 18 weeks to predict a response at the $54^{th}$ week. In conclusion, even though necessarily a bit subjective, it seems that 3 measurements leads to reasonably good quality, while reducing the study time to a third.

For the schizophrenia experiment, first, to stabilize the variance, a linear transformation of the outcome and a non-linear transformation of time, taking the form $Y_{ij} = -3.5675 + 0.0484 \cdot \text{PANSS}_{ij}$ and $t_{j,\text{new}} = e^{-t_j/4}$, respectively, were applied. It follows that, with increasing weight assigned to precision: the first one; the first and the second; or all four time points were required to optimally predict the final measurement. In this case, with similar logic as in the previous case study, it appears that two measurements provides reasonable results, while leading to a 50% study-time reduction.

In line with intuition, in both cases, the number of time points required also changes with increasing $R$. Setting $R = 0$ corresponds to assuming that patients are recruited at no cost or when interest is solely with the cost per additional measurement occasion.

To accommodate the waiting time in the decision making process, we also studied the optimal number of time points based on the modified cost functions (15) and (16). Results can be found in Table 7 for schizophrenia and Table 8 for opthalmology. The modified functions lead to the same results when $R = 0$, but, as $R$ increases, the modified cost functions are more prudent and tend to select less time points.

# 8 Discussion

Our simulation study involved varying numbers of trials and subjects within trials. Unlike conventional surrogate marker validation, which involves two separate outcomes where one is used as a potential surrogate for the other, here we have studied a scenario where there is a single outcome only, measured repeatedly over time. The objective was to assess the performance of accumulated measures of an equally spaced longitudinal sequence as a possible surrogate for a final outcome and to determine the optimal number of repeated measures required to adequately attain 'good' surrogacy. The individual-level surrogacy was assessed using the canonical correlation approach, introduced by Alonso *et al* (2004) and discussed in Section 3. The determination of the optimal number of measurements requires striking a balance between precision and cost of incorporating a long sequence of repeated measures. To this end, an objective function has been utilized. The objective function has two parts, which takes care of the cost and precision components. The im-

Table 6: *Case study in opthalmology. Results for the optimal number of measurements based on cost function (14). ($w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements;$R = C_1/C_2$ be the cost ratio ; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy.)*

| $VRF_{\text{ind}}(K-1) = 0.91$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $w_1$ | $R$ | $m_o$ | $VRF_{\text{ind}}$ | $w_1$ | $R$ | $m_o$ | $VRF_{\text{ind}}$ |
| 0.1 | 0 | 1 | 0.18 | 0.1 | 4 | 1 | 0.18 |
| 0.3 | 0 | 1 | 0.18 | 0.3 | 4 | 1 | 0.18 |
| 0.4 | 0 | 2 | 0.34 | 0.4 | 4 | 3 | 0.45 |
| 0.5 | 0 | 3 | 0.45 | 0.5 | 4 | 8 | 0.85 |
| 0.7 | 0 | 9 | 0.91 | 0.7 | 4 | 9 | 0.91 |
| 0.1 | 1 | 1 | 0.18 | 0.1 | 6 | 1 | 0.18 |
| 0.3 | 1 | 1 | 0.18 | 0.3 | 6 | 2 | 0.34 |
| 0.4 | 1 | 2 | 0.34 | 0.4 | 6 | 3 | 0.45 |
| 0.5 | 1 | 3 | 0.45 | 0.5 | 6 | 8 | 0.85 |
| 0.7 | 1 | 9 | 0.91 | 0.7 | 6 | 9 | 0.91 |
| 0.1 | 2 | 1 | 0.18 | | | | |
| 0.3 | 2 | 1 | 0.18 | | | | |
| 0.4 | 2 | 2 | 0.34 | | | | |
| 0.5 | 2 | 3 | 0.45 | | | | |
| 0.7 | 2 | 9 | 0.91 | | | | |

portance of both components is gauged through the use of weights. Whenever it is felt that the importance of precision outweighs cost, more weight will be assigned to the precision part and vice versa.

The objective function can be modified to accommodate other possible sources of cost. One such cost is the cost of waiting time. This can be incorporated through a third component which accounts for the time lag between the start of the study and the optimal time point. This calls for assigning three possible weights, corresponding to financial cost, time cost, and precision cost, respectively. Similarly, when it is deemed better to detect a condition early rather than late. A possible extension of our work would be to incorporate the cost of a failure to detect the condition early, when treatments are more effective or when a change to an alternative therapy may be more beneficial than when such a switch is effectuated at a later stage.

Table 7: *Case study in schizophrenia. Results for the optimal number of measurements based on cost function (14) and modified cost function (15). ($w_1$: weight assigned to the precision part of the objective function; $m_o$: optimal number of measurements; $R = C_1/C_2$ be the cost ratio ; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy.)*

| $VRF_{\text{ind}}(K-1) = 0.85$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Cost function (14) | | | | Cost function (15) | | | | | |
| $w_1$ | $R$ | $m_o$ | $VRF_{\text{ind}}(m)$ | $w_1$ | $w_2$ | $w_3$ | $R$ | $m_o$ | $VRF_{\text{ind}}(m)$ |
| 0.1 | 0 | 1 | 0.20 | 0.1 | 0.1 | 0.8 | 0 | 1 | 0.20 |
| 0.3 | 0 | 1 | 0.20 | 0.3 | 0.1 | 0.6 | 0 | 1 | 0.20 |
| 0.5 | 0 | 2 | 0.59 | 0.5 | 0.1 | 0.4 | 0 | 2 | 0.59 |
| 0.7 | 0 | 4 | 0.85 | 0.7 | 0.1 | 0.2 | 0 | 4 | 0.85 |
| 0.1 | 1 | 1 | 0.20 | 0.1 | 0.1 | 0.8 | 1 | 1 | 0.20 |
| 0.3 | 1 | 2 | 0.59 | 0.3 | 0.1 | 0.6 | 1 | 1 | 0.20 |
| 0.5 | 1 | 2 | 0.59 | 0.5 | 0.1 | 0.4 | 1 | 2 | 0.59 |
| 0.7 | 1 | 4 | 0.85 | 0.7 | 0.1 | 0.2 | 1 | 4 | 0.85 |
| 0.1 | 2 | 1 | 0.20 | 0.1 | 0.1 | 0.8 | 2 | 1 | 0.20 |
| 0.3 | 2 | 2 | 0.59 | 0.3 | 0.1 | 0.6 | 2 | 1 | 0.20 |
| 0.5 | 2 | 2 | 0.59 | 0.5 | 0.1 | 0.4 | 2 | 2 | 0.59 |
| 0.7 | 2 | 4 | 0.85 | 0.7 | 0.1 | 0.2 | 2 | 4 | 0.85 |
| 0.1 | 4 | 1 | 0.20 | 0.1 | 0.1 | 0.8 | 4 | 1 | 0.20 |
| 0.3 | 4 | 2 | 0.59 | 0.3 | 0.1 | 0.6 | 4 | 1 | 0.20 |
| 0.5 | 4 | 4 | 0.85 | 0.5 | 0.1 | 0.4 | 4 | 2 | 0.59 |
| 0.7 | 4 | 4 | 0.85 | 0.7 | 0.1 | 0.2 | 4 | 4 | 0.85 |
| 0.1 | 6 | 1 | 0.20 | 0.1 | 0.1 | 0.8 | 6 | 1 | 0.20 |
| 0.3 | 6 | 2 | 0.59 | 0.3 | 0.1 | 0.6 | 6 | 1 | 0.20 |
| 0.5 | 6 | 4 | 0.85 | 0.5 | 0.1 | 0.4 | 6 | 2 | 0.59 |
| 0.7 | 6 | 4 | 0.85 | 0.7 | 0.1 | 0.2 | 6 | 4 | 0.85 |

The results of the simulation study for two data-generation schemes, based on CS and AR(1), respectively, have revealed that, depending on the correlation structure of the data and the weights assigned, the first few repeated measures or the entire $K-1$ sequence might be needed to adequately predict the outcome at the last time point. Assuming that the outcome has an AR(1) structure, we showed theoretically and via simulations that either only the first measurement or the entire $K-1$ sequence is required to predict the true endpoint, depending on the weights chosen and the level of the AR(1) correlation. This is a very interesting characteristic of the first-order auto-regressive

Table 8: *Case study in opthalmology. Results for the optimal number of measurements based on modified cost function (15) and (16); ( $w_1$-$w_3$): weights assigned to the precision, financial cost and waiting time parts of the objective function; $m_o$: optimal number of measurements; $R = C_1/C_2$ be the cost ratio; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $f = 100$: percentage of datasets resulting in $m_o$, in all cases.)*

| Weights | | | Cost Ratios | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $R=0$ | | $R=1$ | | $R=2$ | | $R=4$ | | $R=6$ | |
| $w_1$ | $w_2$ | $w_3$ | $m_o$ | $VRF_{\text{ind}}$ | $m_o$ | $VRF_{\text{ind}}$ | $m_o$ | $VRF_{\text{ind}}$ | $m_o$ | $VRF_{\text{ind}}$ | $m_o$ | $VRF_{\text{ind}}$ |
| Modified cost function (15) | | | | | | | | | | | | |
| 0.1 | 0.1 | 0.8 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.1 | 0.6 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.1 | 0.5 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.1 | 0.4 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 |
| 0.7 | 0.1 | 0.2 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 |
| 0.1 | 0.2 | 0.7 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.2 | 0.5 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.2 | 0.4 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.2 | 0.3 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 |
| 0.6 | 0.2 | 0.2 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 | 9 | 0.91 | 9 | 0.91 |
| 0.1 | 0.3 | 0.6 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.3 | 0.4 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.3 | 0.3 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.3 | 0.2 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 8 | 0.85 |
| 0.6 | 0.3 | 0.1 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 | 9 | 0.91 | 9 | 0.91 |
| Modified cost function (16) | | | | | | | | | | | | |
| 0.1 | 0.1 | 0.8 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.1 | 0.6 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.1 | 0.5 | 1 | 0.18 | 1 | 0.18 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.1 | 0.4 | 2 | 0.34 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 |
| 0.7 | 0.1 | 0.2 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 | 9 | 0.91 |
| 0.1 | 0.2 | 0.7 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.2 | 0.5 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.2 | 0.4 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.2 | 0.3 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 |
| 0.6 | 0.2 | 0.2 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 |
| 0.1 | 0.3 | 0.6 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.3 | 0.3 | 0.4 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 | 1 | 0.18 |
| 0.4 | 0.3 | 0.3 | 2 | 0.18 | 2 | 0.18 | 2 | 0.34 | 2 | 0.34 | 2 | 0.34 |
| 0.5 | 0.3 | 0.2 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 | 3 | 0.45 |
| 0.6 | 0.3 | 0.1 | 8 | 0.85 | 8 | 0.85 | 8 | 0.85 | 9 | 0.91 | 9 | 0.91 |

structure. Our results illustrate that here no balance between precision and cost is possible, because the $CPR$ always leads to the two extreme situations. If precision is the driving requirement, then the entire $K-1$ subsequence is the best option, whereas if cost if the impelling factor then the surrogate should never contain more than a single observation. In such a situation, the best strategy will be to use only one measurement, located somewhere in the interval $(1, k-1)$. Obviously if the observation is taken at the end of the sequence, more predictive power will be achieved but a longer waiting time will also be needed. Arguably, a decision should then be taken based on other field related factors and the opinion of the experts in the area will be important. Moreover, at most six measurements, about 60% of the entire sequence, are required to adequately predict the final measurement if the outcome has a CS or a Toeplitz structure, or a general structure with slowly decaying correlation between repeated measures.

Based on these findings, it seems promising to use the proposed approach to balance between cost and precision in the process of evaluating the performance of a few repeated measures taken early as possible surrogates to adequately predict the outcome and/or treatment effect of the final measure.

Our simulation study, while relatively broad, is intrinsically limited, as is the case for every simulation study. A number of extensions could be considered. First, while the first-order autoregressive structure applies to equally spaced measures only, this is not the case for the compound symmetry and unstructured covariances. In principle, further structures for unbalanced data, such as general special functions, as available in the SAS System, could be considered. Second, our derivations crucially rely on the continuous nature of the outcome, and hence on the linearity of the expressions involved, enabling the derivation of explicit expressions. Should the outcome be non-Gaussian, then relevant model choices are generalized estimating equations (GEE, Liang and Zeger 1986) or generalized linear mixed models (GLMM, Breslow and Clayton 1993), for example. A review of this and additional methodology is provided in Molenberghs and Verbeke (2005). Such models, however, raise a number of complexities. The presence of a mean-variance link and the non-linear nature of the link function defeats the derivation of explicit analytical expressions like in the continuous case. Of course, one might make progress through the use of approximate expressions, or by way of Monte-Carlo-based evaluations. These are just two examples of how extensions could be considered.

The analyses in this paper can be carried out using commonly available software such as SAS. A SAS macro can be obtained from the authors' web pages.

## Acknowledgments

Financial support from the IAP research network #P6/03 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The authors are indebted to (OSI) Eyetech Pharmaceuticals for the kind permission to use their clinical-trial data.

## 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., Geys, H., Molenberghs, G., Kenward, M., and Vangeneugden, T. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurments: Canonical correlation approach. *Biometrics*, **60**, 845–853.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

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

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

Galecki, A. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: Theory and Methods*, **23**, 3105–3119.

Graybill, F.A. (1983). *Matrices with Applications in Statistics ( 2nd ed.)* Belmont, California: Wadsworth.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin* **13**, 261–276.

Kay, S.R., Opler, L.A., and Lindenmayer, J.P. (1988). Reliability and validity of the Positive and Negative Syndrome Scale for Schizophrenics. *Psychiatric Research* **23**, 99–110.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

Tilahun, A., Assam, P., Alonso, A., and Molenberghs, G. (2007). Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Computational Statistics and Data Analysis*, **51**, 4152–4163.

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

Winkens, B., Schouten, H.J.A, van Breukelen, G.J.P., and Berger, M.P.F. (2005). Optimal time-points in clinical trials with linearly divergent treatment effects. *Statistics in Medicine*, **24**, 3743–3756.