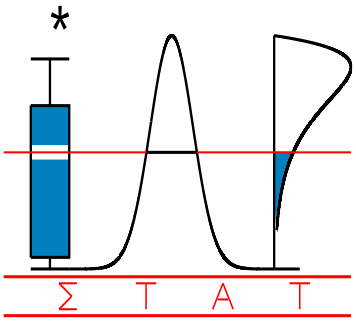
T E C H N I C A L

R E P O R T

06100

ALTERNATIVE METHODS TO EVALUATE TRIAL LEVEL SURROGACY

CORTINAS ABRAHANTES, J., SHKEDY Z. and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Alternative Methods to Evaluate Trial Level Surrogacy

José Cortiñas Abrahantes, Ziv Shkedy, Geert Molenberghs

Center for Statistics, Hasselt University, Campus Diepenbeek,

B3590 Diepenbeek, Belgium

Abstract

The evaluation and validation of surrogate endpoints have been extensively studied in the last decade. Prentice (1989) and Freedman, Graubard and Schatzkin (1992) laid the foundations for the evaluation of surrogate endpoints in randomized clinical trials. Buyse et al. (2000) proposed a meta-analytic methodology, producing different methods for different settings, which was further studied by Alonso and Molenberghs (2006), in their unifying approach based on information theory. In this paper, we propose alternative procedures to evaluate the so-called trial-level surrogacy and a correction based on cross-validation ideas. We then apply the various strategies to data from three clinical studies: Pharmacological Therapy for Macular Degeneration Study Group (1977), Four Meta-analyses of 28 Clinical Trials in Advance Colorectal Cancer (1996) and a Meta-analysis of Five Clinical Trials in Schizophrenia (1996). The results obtained indicate that using random forest or bagging models produces larger estimated values for the surrogacy measure, which were in general stabler and the confidence interval narrower than the other methods employed to estimate this quantity.

Some Keywords: Linear mixed model; Macular degeneration; Meta-analytic approach; Oncology; Random effects; Surrogate endpoint.

1 Introduction

Prentice (1989) and Freedman, Graubard and Schatzkin (1992) laid the foundations for the evaluation of surrogate endpoints in randomized clinical studies. Prentice proposed a definition as well as a set of operational criteria, while Freedman, Graubard and Schatzkin (1992) supplemented these criteria with a quantity called *proportion explained* (PE), which was meant to indicate the proportion of the treatment effect mediated by the surrogate. Later, Buyse and Molenberghs (1998) proposed to use instead the *relative effect* (RE), linking the effect of treatment on both endpoints and a second measure called individual-level which measure the agreement between both endpoints, after adjusting for the effect of treatment (*adjusted association*). This suffers from to untestable assumptions and low statistical power. In order to overcome these problems, several authors (Daniels and Hughes, 1997; Buyse et al., 2000; Gail et al., 2000) have proposed methods that combined evidence from several clinical trials, such as

in a meta-analysis, rather than from a single study. To this end, a bivariate hierarchical model was formulated, accommodating the surrogate and true endpoints in a multi-trial setting. In Buyse et al. (2000), the *adjusted association* carries over when data are available on several randomized trials, while the RE needed to be extended to what is now a called trial-level measure of agreement between the effects of treatment on both endpoints. This modifies the relative effect and the adjusted association to become a trial-level R^2 and an individual-level R^2, respectively. Similar routes have been followed by Daniels and Hughes (1997) and Gail et al. (2000).

While the proposal is elegant, it suffers from several drawbacks. First, separate developments are necessary for various types of endpoints. Buyse et al. (2000) considered normally distributed endpoints. An overview of corresponding methods for binary, time-to-event, and longitudinal endpoints can be found in Burzykowski, Molenberghs and Buyse (2005). The main issue is that, especially the individual-level surrogacy, is captured through a disparate range of measures. Second, estimation within a hierarchically formulated model framework can be challenges, for which simplified model strategies had to be developed (Tibaldi *et al.*, 2003), generally based on replacing a hierarchical analysis by a two-stage alternative, where first trials are analysed separately, after which relevant summary measures are combined into a single analysis. Finally, even when the hierarchical model is within reach, the resulting point estimates and precision measures may be less than reliable.

Regarding the first concern, also Alonso and Molenberghs (2006) discussed in their paper the limitations of the meta-analytic methodology, resulting in the aforementioned collection of definitions for the individual-level surrogacy measure, depending on the type of endpoint. To compound the issue, these measures are sometimes expressed at a latent level, whereas they are explicitly in terms of the observed outcomes in other situations; this clearly compounds the issue. In response to these issues, Alonso and Molenberghs (2006) proposed a unifying approach based on information theory. Fortunately, trial-level surrogacy has always been measured using the determination coefficient that results from the regression between the effect of treatment on the true and the surrogate endpoints.

In the present article, we address the other concerns, regarding the validity of the trial-level surrogacy estimates. Conventionally, estimation is based on fitting a linear mixed-effects model (Verbeke and Molenberghs, 2000) or one of its simplification outlined in Tibaldi *et al.* (2003). The corresponding

standard errors and interval estimates for the trial-level surrogacy generally derive from the delta method. It will be shown here that the so-obtained results can be unreliable or even plain misleading. Therefore, a collection of alternative methods, based on regression trees, random forests, and support vector machines, combined with bootstrap-based confidence interval and, should one wish, in conjunction with a cross-validation based correction, will be proposed and applied. The corresponding computer code is made available through the authors' web pages.

In Section 2, three motivating case studies are introduced, together with results from the original analyses. The two-stage model, to be used throughout the paper, is presented in Section 3. The proposed methods are described in Section 4. Section 5 present the results of applying these methods to the case studies.

2 Motivating Case Studies

We consider three case studies, covering important and different therapeutic areas. Earlier analyses, to be contrasted with ours, can be found in Burzykowski, Molenberghs and Buyse (2005).

The first one is situated within ophthalmology, the second one is from advanced colorectal cancer, and the final one is a psychiatric study. We will compare our results regarding trial-level surrogacy with those reported in Burzykowski, Molenberghs and Buyse (2005).

2.1 The Age-Related Macular Degeneration Study (ARMD)

These data come from a randomized clinical trial comparing an experimental treatment (interferon-α) to placebo in the treatment of patients with age-related macular degeneration. The aim of the study was to compare placebo and the highest dose of interferon-α. Since we have a single multi-centric trial, i refers to center and j to patient within center. The true endpoint in this study was the change in visual acuity at 12 months after starting the treatment. The surrogate endpoint considered is visual acuity at 6 months.

2.2 Advanced Colorectal Cancer

We consider data from four randomized multicenter trials in colorectal cancer. These constitute the largest source of randomized data available in advanced colorectal cancer. All data were collected and

checked by the Meta-Analysis Group In Cancer between 1990 and 1996 (Corfu-A Group, 1995; Greco et al., 1996) to confirm the benefits of experimental fluoropyrimidine treatments with 5-fluorouracil (5FU) in advanced colorectal cancer. The principal investigators of all trials provided data for every patient, whether eligible or not, and whether properly followed-up or not. Burzykowski, Molenberghs and Buyse (2004) and Burzykowski, Molenberghs and Buyse (2005) provide full details on the trials included the treatments tested, the patient characteristics, and the therapeutic results.

In this study, we compare 5FU plus interferon with 5FU alone. The final endpoint is survival time in years, while the surrogate is progression-free survival time, i.e., the years between the randomization to clinical progression of the disease or death. In agreement with previous analyses, only centers with at least 3 patients on each treatment arm are considered. The data include 48 centers, with a total sample size of 642 patients.

2.3 Clinical Studies in Schizophrenia

The psychiatric studies in schizophrenic patients is based on a meta-analysis containing five trials (Alonso et al., 2002). This is insufficient to apply the meta-analytic methods. Instead, we will used country as a unit of analysis. Note that the choice of units is an important issue, and should be carefully considered (Cortiñas et al., 2004). The true endpoint is Clinician's Global Impression (CGI). This is a 7-grade scale, frequently used by the treating physician to characterize how well a subject is doing. As a surrogate measure, we consider the Positive and Negative Syndrome Scale (PANSS, Kay et al. (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. There are 20 country-units, with the number of patients per unit ranging from 9 to 128.

3 The Two-Stage Approach

Let us introduce a set of notation that will be used throughout the paper. Let $Y_{T_{ij}}$ and $Y_{S_{ij}}$ be random variables denoting the true and the surrogate endpoints for subject $j = 1, \ldots n_i$ in unit $i = 1, \ldots N$. Further, let Z_{ij} denote a binary treatment indicator.

3.1 A Two Stage Meta-analytic Approach

The hierarchical two stage approach proposed by Buyse et al. (2000), based on the two-stage fixed-effects representation is:

$$\begin{cases} Y_{S_{ij}} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, \\ Y_{T_{ij}} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \end{cases} \tag{1}$$

where μ_{S_i} and μ_{T_i} are unit-specific intercepts, α_i and β_i are unit-specific treatment effects on the endpoints in unit i, and $\varepsilon_{S_{ij}}$ and $\varepsilon_{T_{ij}}$ are correlated error terms.

At the second stage, it is assumed that

$$\begin{cases} \mu_{S_i} = \mu_S + m_{S_i}, \\ \mu_{T_i} = \mu_T + m_{T_i}, \\ \alpha_i = \alpha + a_i, \\ \beta_i = \beta + b_i, \end{cases} \tag{2}$$

The authors assumed that the two endpoints were normally distributed and at the second stage the μ_S and μ_T are fixed intercepts, m_{S_i} and m_{T_i} are random intercepts for the unit i, α and β are fixed treatment effects and a_i and b_i are random treatment effects. The vector of random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)^T$, was assumed to be zero-mean normally distributed with variance-covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Sa} & d_{ab} & d_{bb} \end{pmatrix}. \tag{3}$$

Other representations, such as the random-effects representation can be used, in which both steps are combined. In the context of surrogate endpoint validation both approaches typically perform very similarly.

3.2 Trial-Level Surrogacy

We will focus on the evaluation of trial-level surrogacy. The key motivation for validating a surrogate endpoint is the wish to predict the effect of treatment on the true endpoint based on the observed effect of treatment on the surrogate endpoint. Suppose we consider a new trial, $i = 0$ say, for which data are available on the surrogate endpoint but not on the true endpoint. We are interested in the estimated effect of Z on Y_T, given the effect of Z on Y_S for this particular trial. Let us subscript all

quantities pertaining to the particular trial under study with 0. It is easy to show (Buyse et al., 2000) that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \quad (4)$$

$$\mathsf{Var}(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (5)$$

Related to prediction equations (4)–(5), a measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$R^2_{\text{trial (f)}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (6)$$

A good surrogate, *at the trial level*, would have (6) close to 1, which will be associated with a surrogate for whom the variance of $(\beta + b_0|m_{S0}, a_0)$ is zero.

Intuition can be gained by considering the simplified case where the prediction of b_0 is done independently of the random intercept m_{S0}. The coefficient (6) then reduces to

$$R^2_{\text{trial (r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \quad (7)$$

This formula is useful when the full random-effects model is hard to fit but a reduced version, excluding random intercepts, is easier to reach convergence. Note that all methods are essentially rooted in the concept of regression the true endpoint on the surrogate endpoints, perhaps using auxiliary information from the intercept (background effect) and properly taking the information into account, ideally through a two-stage approach.

In practice, in many situation we found that either the surrogate, or the true endpoint, or both are of a non-Gaussian type. In the next sections we will lay the foundations for the alternative approaches. The starting point is method that have been applied to the case studies before.

3.3 A Copula Modeling Approach for Categorical Surrogate and Survival True Endpoint

Burzykowski, Molenberghs and Buyse (2004) extended the methodology proposed by Buyse et al. (2000), when the surrogate is a categorical variable with K ordered categories and the true endpoint is a failure-time random variable. The authors replace the first stage model by a bivariate copula model for the

true and a latent continuous variable underlying the surrogate endpoint. The full bivariate model corresponding to (1) assumed that the joint cumulative distribution of $Y_{T_{ij}}$ (the true endpoint) and $Y_{\tilde{S}_{ij}}$ (the surrogate endpoint) given $Z_{ij} = z$, is generated by a one-parameter copula function C_θ:

$$F_{Y_{\tilde{S}_{ij}},Y_{T_{ij}}}(y_T, y_S; z) = C_\theta \left[F_{Y_{\tilde{S}_{ij}}}(y_S; z), F_{Y_{T_{ij}}}(y_T; z), \theta \right], \tag{8}$$

where $C_\theta[.,.]$ is a distribution function on $[0,1]^2$ with $\theta \epsilon \Re$, describing the association between $Y_{\tilde{S}_{ij}}$ and $Y_{T_{ij}}$. An attractive feature of this model is that the marginal models (proportional odds and proportional hazards models) and the association model can be selected without constraining each other.

At the first stage Burzykowski, Molenberghs and Buyse (2004) proposed using the maximum likelihood estimates of the parameters of model (8), assuming a fixed-effects representation. In (8), trial-specific treatment effects α_i and β_i on the surrogate and the true endpoint were estimated. At the second stage, the authors proposed to evaluate the trial-level surrogacy using the determination coefficient from the linear regression of β_i on α_i.

3.4 A Joint Modelling Approach for Longitudinal Surrogate and True Endpoints

Alonso et al. (2003) extended the methodology proposed by Buyse et al. (2000) to the case where both endpoints are longitudinal. This setting poses important challenges in terms of, first, finding a model that can accommodate such multivariate structures of the data and finally new measures that allow us to evaluate surrogacy when both endpoints are of this type.

Assume further that ξ_{ijk} is the time corresponding to the kth occasion ($k = 1, \ldots, p_i$) when subject j in trial i was measured. Following the ideas of Galecki (1994), Alonso et al. (2003) proposed a specific joint model at the first stage for both responses:

$$\begin{cases} Y_{S_{ijk}} = \mu_{S_i} + \alpha_i Z_{ij} + g_{T_{ij}}(\xi_{ijk}) + \varepsilon_{S_{ij}}, \\ Y_{T_{ijk}} = \mu_{T_i} + \beta_i Z_{ij} + g_{S_{ij}}(\xi_{ijk}) + \varepsilon_{T_{ij}}, \end{cases} \tag{9}$$

where μ_{S_i} and μ_{T_i} are as for model (1) unit-specific intercepts, α_i and β_i are unit-specific effects of treatment Z_{ij} on the two endpoints and $g_{S_{ij}}$ and $g_{T_{ij}}$ are trial and subject-specific time functions. Note that, even though in practice $Y_{S_{ij}}$ and $Y_{T_{ij}}$ are frequently measured at the same time points, model (9) does not preclude the more general case. The random vectors associated to the error for both endpoints

are assumed to jointly follow a mean-zero multivariate normal distribution with variance-covariance matrix

$$\Sigma_i = \begin{pmatrix} \sigma_{SS_i} & d_{ST_i} \\ d_{ST_i} & d_{TT_i} \end{pmatrix} \otimes R_i, \tag{10}$$

where R_i reflects a general correlation matrix for the repeated measurements. More details can be found in Alonso et al. (2003). If treatment effect is assumed constant over time, then the R^2_{trial} measure proposed by Buyse et al. (2000) would be used to evaluate surrogacy at the trial level.

3.5 The Original Analyses of the Case Studies

For the ARMD trial, Buyse et al. (2000) experienced problems in fitting the full random-effects models. Therefore, they entertained a (unweighted) fixed-effects approach instead, based on ideas of Tibaldi *et al.* (2003). This produced a moderate trial-level surrogacy: $R^2_{\text{trial (f)}} = 0.692$. The standard errors were calculated by means of a straightforward application of the delta method, based on deriving the variance of R^2 from its Fisher's z transform variance and then producing a confidence interval of $[0.518; 0.866]$. Equipped with our newly proposed tools, we will revisit this conclusion in Section 5.1.

For the advanced colorectal cancer case, Burzykowski, Molenberghs and Buyse (2004) considered a so-called landmark time of 3 months, which produced a trial-level surrogacy of $R^2_{\text{trial (r)}} = 0.15$ with 95% confidence interval $[0; 0.41]$. This clearly is absolutely too low to even consider moving forward with this particular candidate for surrogacy. However, we will shed a different light on this case in Section 5.2.

Considering the schizophrenia trials, a two-stage model was fitted to these data, incorporating a linear trend over time, found to be the best fitting, parsimonious model, as a result of a model-building exercise that set out by allowing for random splines (Verbyla et al., 1999; Alonso et al., 2004). The trial-level surrogacy obtained for this case study was: $R^2_{\text{trial (f)}} = 0.820$ with 95% confidence interval $[0.611; 0.920]$. Details can be found in Alonso et al. (2004). We will return to this study in Section 5.3.

4 Alternative Procedures to Evaluate Trial-Level Surrogacy

In this section, we present other techniques that can be used to obtain the trial-level surrogacy using the two-stage approach proposed by Buyse et al. (2000). In their approach, they proposed to use the determination coefficient that comes out from the regression between the effects on treatment from both endpoints ($\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$). There are two major points that

need to be carefully revised in this definition of trial level surrogacy. First, the estimates that we are obtaining to evaluate trial level surrogacy are over-optimistic, in the sense that the data used to build the model are also used to evaluate the fit of the final model. To appropriately account for this issue, we will use a k-fold cross-validation method to obtain a fair estimate of the trial-level surrogacy measure. All alternative methods used to estimate trial level surrogacy will be subjected to a 10-fold cross-validation correction. Second, in general, the relation between these effects does not necessarily need to be linear. This is why we propose the use of more flexible regression techniques instead, which allow for a more general functional relation between the effects of treatments on both endpoints. We will focus on regression trees (Breiman et al., 1984), bagging algorithms (Breiman, 1996a,b), random forest (Breiman, 2001) and support vector regression (Vapnik, 1995).

Delta-method confidence intervals can be misleading, as we will show by comparison. This motivates the use of a bootstrap alternative, which is generally applicable. Also, specifically for the linear-regression case, Ding's method can be used (Ding, 1996), based on reporting the 2.5 and 97.5 quantiles of the cumulative distribution function of R^2.

We used the R statistical computing environment (Ihaka and Gentleman, 1996) and the R packages RPART version 3.1-27 (Therneau and Atkinson, 1997), randomForest version 4.5-15 (Liaw and Wiener, 2002) and the interface libsvm from e1071 version 1.5-12 (Meyer, 2001).

4.1 Regression Tree Analysis

The regression tree methodology is a very well-known and widely used technique (Therneau and Atkinson, 1997). Unlike classical regression techniques for which the relationship between the response and predictors is pre-specified, such as linear or quadratic, and the test is performed to confirm or reject the relationship, regression tree analysis (RTA) assumes no such relationship. It is primarily a method for constructing a set of decision rules on the predictor variables (Breiman et al., 1984; Verbyla, 1987). The rules are constructed by recursively partitioning the data into successively smaller groups with binary splits based on a single predictor variable. Splits for all of the predictors are examined by an exhaustive search procedure and the best split is chosen. For regression trees, the selected split is the one that maximizes the homogeneity of the two resulting groups with respect to the response variable, the split that maximizes the between-group sum of squares, as in analysis of variance, although other options

may be available. The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response. The procedure initially grows maximal trees and then uses techniques such as cross-validation to prune the overfitted tree to an optimal size (Therneau and Atkinson, 1997). RTA has clear advantages over classical statistical methods. It is effective in uncovering structure in data with hierarchical or non-additive variables. Because no a priori assumptions are made about the nature of the relationships among the response and predictor variables, RTA allows for the possibility of interactions and non-linearity among variables (Moore, Lees and Davey, 1991). Details about the methods can be found in Therneau and Atkinson (1997).

The trial-level surrogacy measure that will be employed when regression tree analysis is used is given by the relative reduction in deviance of the final tree.

$$RD_{tree} = \frac{D(\beta) - D(\beta \mid \alpha)}{D(\beta)}, \tag{11}$$

where $D(\beta)$ denotes the deviance or total variability of the effects of treatments for the true endpoint; it is given by the following expression:

$$D(\beta) = \sum_{i=1}^{N} (\beta_i - \bar{\beta})^2. \tag{12}$$

Furthermore, $D(\beta \mid \alpha)$ denotes the deviance of the final pruned tree when the information of the effects of treatments for the surrogate endpoint is used. Assuming that we have m final nodes ($M_1, M_2, \ldots, M_m$), then $D(\beta \mid \alpha)$ can be calculated as follows:

$$D(\beta \mid \alpha) = \sum_{h=1}^{m} \left(\sum_{\beta_i \in M_h} (\beta_i - \overline{\beta_{M_h}})^2 \right), \tag{13}$$

where $\overline{\beta_{M_h}}$ is the mean of the effects of treatment on the true endpoint in terminal node M_h.

4.2 Bagging Regression Trees

Bagging, a contraction of '**b**ootstrap **agg**regat**ing**', is a technique proposed by Breiman (1996a,b) that can be used with many regression methods so as to reduce the variance associated with prediction, thereby improving the prediction process. It is a relatively simple idea: many bootstrap samples are drawn from the available data, some prediction method is applied to each bootstrap sample, and then the results are combined, by averaging for regression, to obtain the overall prediction, with the variance

being reduced due to the averaging. It can be used to improve both the stability and predictive power of regression trees, but its use is not restricted to improving tree-based predictions. It is a general technique that can be applied in a wide variety of settings to improve predictions. Details about the method can be found in Breiman (1996a).

The trial-level surrogacy measure will be the median of the list of relative reduction in deviance RD_{tree} of each tree constructed for each bootstrap sample. In our case 1000 bootstrap samples were constructed.

4.3 Random Forests (RF)

The random forest method (Breiman, 2001) is a supervised learning algorithm that has previously been successfully applied to many different types of studies. A random forest is an ensemble of many identically distributed trees generated from bootstrap samples of the original data. Each tree is constructed via a regression tree algorithm. The simplest random forest with random features is formed by selecting randomly, at each node, a small group of input variables to split on. The size of the group is fixed throughout the process of growing the forest. Each tree is grown by using the CART methodology without pruning.

Some features of random forest worth highlighting are: (1) it is an excellent classifier, comparable in accuracy to support vector machines; (2) it generates an internal unbiased estimate of the generalization error as the forest building progresses; (3) it has a method for balancing error in unbalanced population data sets; (4) it computes proximities between pairs of cases that can be used in clustering, locating outliers, or by scaling, give useful views of the data; (5) it is well known that random forests avoid overfitting and usually have better performance than regression trees. Details about random forest can be found in (Breiman, 2001). The trial-level measure of surrogacy will be computed similarly to the case in which bagging methods were used.

4.4 Support Vector Machine (SVM)

The term *support vector machines* (SVM) refers to a family of learning algorithms which is nowadays considered as one of the most efficient methods in throughout a variety of applications. SVM is a supervised learning technique for classification and regression. The SVM algorithm is a non-linear generalization of the so-called Generalized Portrait Algorithm developed in the sixties by Vapnik and

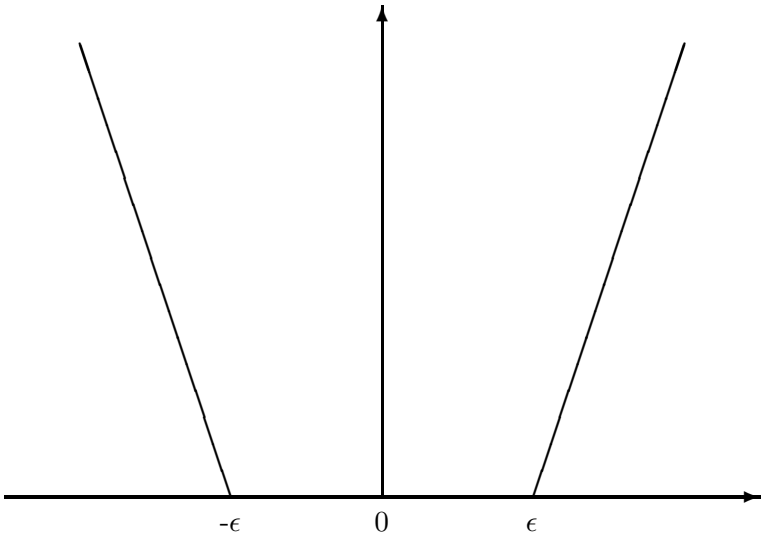


Figure 1: *A piecewise linear ϵ-insensitive loss function.*

Lerner (1963) and Vapnik and Chervonenkis (1964), but the first practical implementation was only published in the early nineties. Ever since, the popularity of the method has been growing among the machine learning and statistical communities.

SVM can also be applied to regression problems by the introduction of an alternative loss function, (Smola, 1996). The loss function must be modified to include a distance measure. SVM regressions uses the ϵ-insensitive loss function show in Figure 1.

If the deviation between the predicted and actual values is less than ϵ, then the regression function is considered good, which can be mathematically expressed as: $-\epsilon \leq \omega \cdot \alpha_i - b - \beta_i \geq \epsilon$.

From a geometric point of view, it can be seen as a band of size 2ϵ around the hypothesis function and any point outside this band is considered as a training error. Suppose the data can be explained by a linear model, the goal is to find a fitting hyperplane $\langle w, \alpha_i \rangle + b = 0$. Formally, we need to minimize $\| \omega \|^2/2$ subject to the following constraints:

$$\begin{aligned} \beta_i - \langle w, \alpha_i \rangle - b &\leq \epsilon, \\ \langle w, \alpha_i \rangle + b - \beta_i &\geq \epsilon. \end{aligned}$$

To account for training errors and the possibility of handling non-linearity we can map the input data α_i into a, possibly higher dimensional, so-called feature space ($\phi(\alpha_i)$) and introduce some weights to

our optimization problem, which now becomes:

$$\min \frac{\| \omega \|^2}{2} + C \cdot \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i),$$

subject to the following constraints:

$$\begin{aligned} \beta_i - \langle w, \phi(\alpha_i) \rangle - b &\leq \epsilon + \xi_i, \\ \langle w, \phi(\alpha_i) \rangle + b - \beta_i &\geq \epsilon + \hat{\xi}_i, \\ \xi_i, \hat{\xi}_i &\geq 0. \end{aligned}$$

We then need to solve an optimization problem with some constraints. It turns out that in most cases it can be solved more easily in its dual formulation. Moreover, the dual formulation provides the key for extending SVM to nonlinear functions. Hence we will use a standard dualization method utilizing Lagrange multipliers, as described in Fletcher (1989). More details can be found in Vapnik (1995).

The trial level surrogacy measure can be computed using the ratio between the portion of the variability not explained by the model and the total variability of the effects of the treatment in the true endpoint:

$$RD_{SVMR} = \frac{D(\beta) - D_{SVMR}(\beta \mid \alpha)}{D(\beta)}.$$

Here, $D(\beta)$ can be calculated using (12), and $D_{SVMR}(\beta \mid \alpha)$ is the sum of the squares of the differences between the actual value (β_i) and their estimated value obtained when the SVM regression model is employed.

5 Results

The trial level surrogacy measure was calculated for each of the three case studies using the five different approaches, in combination with or without 10-fold cross-validation correction. Parameter estimates and standard errors are summarized in Table 1.

5.1 Age Related Macular Degeneration Study (ARMD)

Figure 2(a) shows the scatterplot of the treatment effect estimated for each center, for both endpoints.

It is clear from examining Table 1 that applying cross-validation produces a downward correction across the method applied. From the density plots, presented in Figures 2(b), we discern an asymmetric shape

Table 1: *Estimates [95% confidence intervals] for trial-level surrogacy in the age-related macular degeneration (ARMD), advanced colorectal, and schizophrenia datasets based on the conventional linear model, regression trees, bagging of regression trees, random forests, and support vector regression. Calculations are done without and with cross-validation. Confidence interval are based on the bootstrap, except for the linear model, in which case additionally the delta method and Ding's method is used.*

Method	ARMD	Adv. Colorectal	Schizophrenia
Without cross-validation			
Linear model	0.685 [0.477;0.841]	0.151 [0.014;0.461]	0.805 [0.602;0.900]
Linear model (delta)	0.685 [0.507;0.863]	0.151 [-0.113;0.415]	0.805 [0.638;0.971]
Linear model (Ding)	0.685 [0.463;0.822]	0.151 [0.000;0.438]	0.805 [0.556;0.915]
Regression tree	0.744 [0.604;0.921]	0.472 [0.305;0.851]	0.698 [0.628;0.967]
Bagged regr. tree	0.839 [0.763;0.961]	0.567 [0.441;0.734]	0.811 [0.633;0.936]
Random forest	0.884 [0.842;0.971]	0.623 [0.454;0.833]	0.866 [0.706;0.937]
Support vector machine	0.830 [0.633;0.950]	0.450 [0.157;0.738]	0.830 [0.625;0.949]
With cross-validation			
Linear model	0.618 [0.381;0.824]	0.003 [0.000;0.297]	0.756 [0.473;0.876]
Linear model (delta)	0.618 [0.413;0.823]	0.003 [-0.041;0.047]	0.756 [0.554;0.958]
Linear model (Ding)	0.618 [0.374;0.780]	0.003 [0.000;0.156]	0.756 [0.469;0.892]
Regression tree	0.620 [0.352;0.854]	0.293 [0.021;0.654]	0.497 [0.264;0.902]
Bagged regr. tree	0.693 [0.574;0.921]	0.279 [0.099;0.654]	0.661 [0.514;0.897]
Random forest	0.712 [0.514;0.914]	0.344 [0.046;0.696]	0.621 [0.408;0.863]
Support vector machine	0.684 [0.223;0.890]	0.294 [0.022;0.630]	0.717 [0.133;0.927]

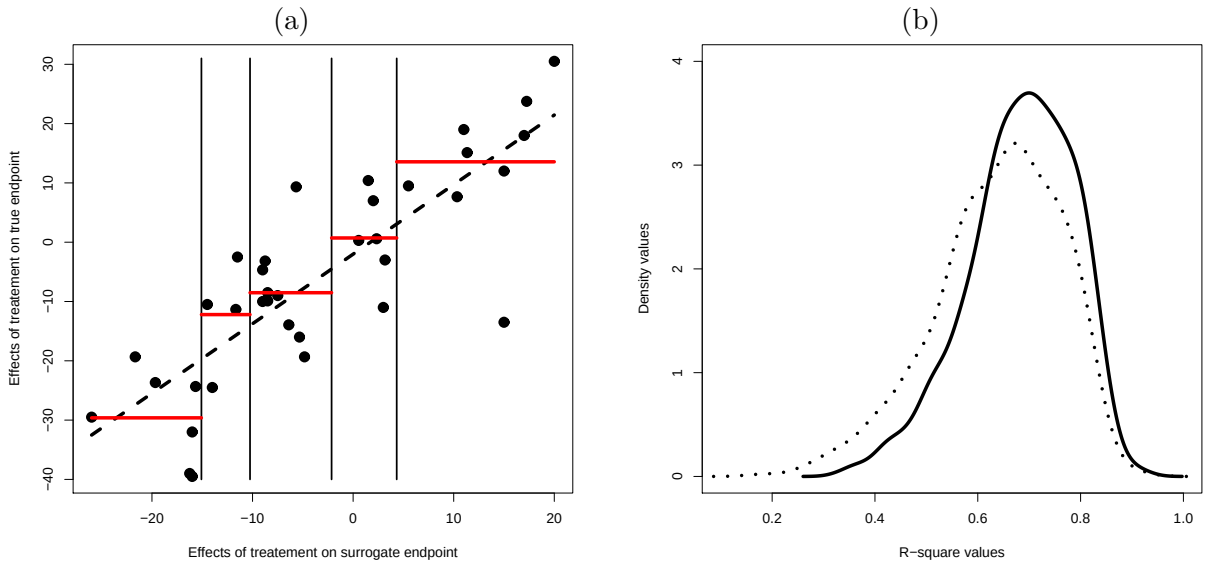


Figure 2: *Age-related macular degeneration study. (a) Scatterplot of the estimated treatment effects for both endpoint in each center, overlaid by predictions of the final pruned tree (solid lines), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate for the ARMD data together with the linear model predictions (dotted line). (b) Estimated density function for the trial level surrogacy, without correction (solid line) and with cross-validation correction (dashed line).*

of the distribution for the trial-level surrogacy measures, whether or not cross-validation is applied. This indicates that a delta interval, by definition symmetric, is less appropriate. The regression tree model was fitted and the fit of the resulting pruned tree, together with the linear model fit is shown in Figure 2(a). Turning to the support vector regression, several kernels and associated parameterizations could be used. Both of these were tuned and the best choice was based on the performance of the model using the sum of the squared residuals with cross-validation, resulting in the radial kernel.

Focusing on the cross-validation outcomes, the point estimates are all reasonable similar, with the random forest based estimate the largest, followed by the bagged regression tree and support vector regression versions. There is considerable difference between the confidence intervals, even when confining attention to the bootstrap based ones. The SVM interval is very wide, with a length of 0.667. Then, there is a middle group, consisting of regression trees (0.502), the linear model (0.440), and random forests (0.400). Finally, the shortest interval is found with bagged regression trees (0.347).

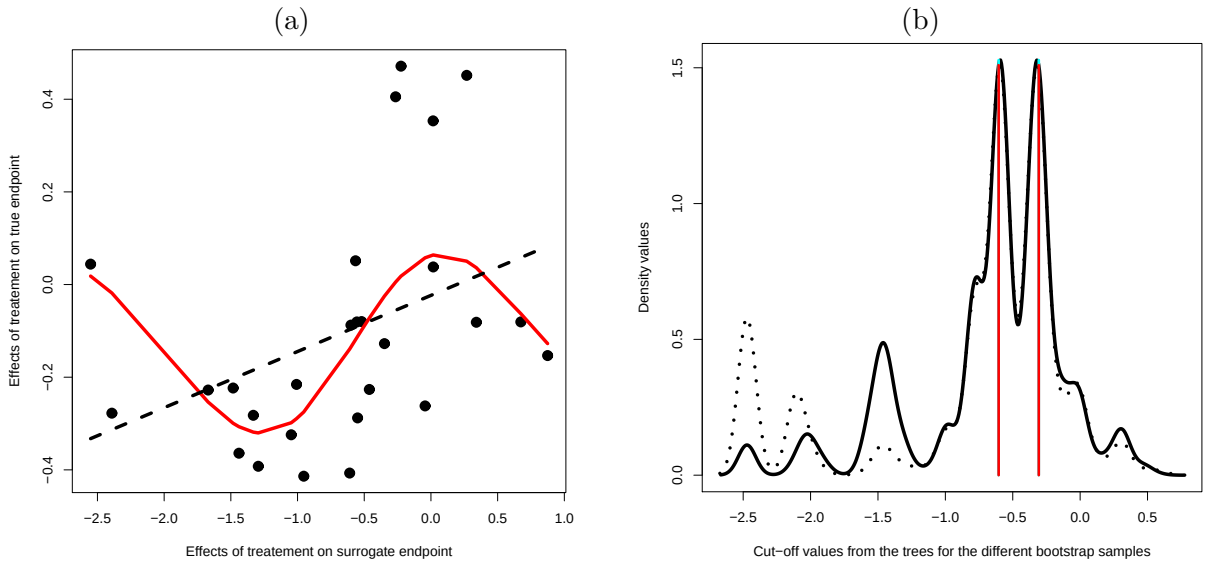


Figure 3: *Advanced Colorectal Cancer. (a) Scatterplot of the estimated treatment effects for both endpoint in each center, overlaid by predictions from support vector regression (solid line), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate, together with the linear model predictions (dashed lines). (b) Estimated density function for the cut-off values used in all 1000 trees and the ones obtained in the final pruned tree (vertical lines). The solid curve refers to bagging; the dotted line represents the random forests.*

5.2 Advanced Colorectal Cancer

We estimated the treatment effect for each unit using the methodology presented in Section 3.3. The scatterplot of the estimated treatment effect for both endpoints in each center is shown in Figure 3(a). As is clear from Table 1, especially using the linear model, and then in particular when cross-validation is employed, the magnitude of the association in this study is much lower than what was observed with the ARMD trial. As a result, the delta interval produces an undesirable negative lower limit in this case. Fortunately, both Ding's method and the bootstrap can be employed to satisfactorily overcome this pitfall. The point estimates for the other methods are all considerable higher, ranging from 0.450 to 0.623 without, and being close to 0.300 with cross-validation.

Figure 3(b) displays the density of cut-off values obtained when random forest is used. It can be seen here that both bagging and random forests report the cut off values obtained in the final pruned tree as the more likely to happen, but they do report other possible cut-offs, which differ between both methods.

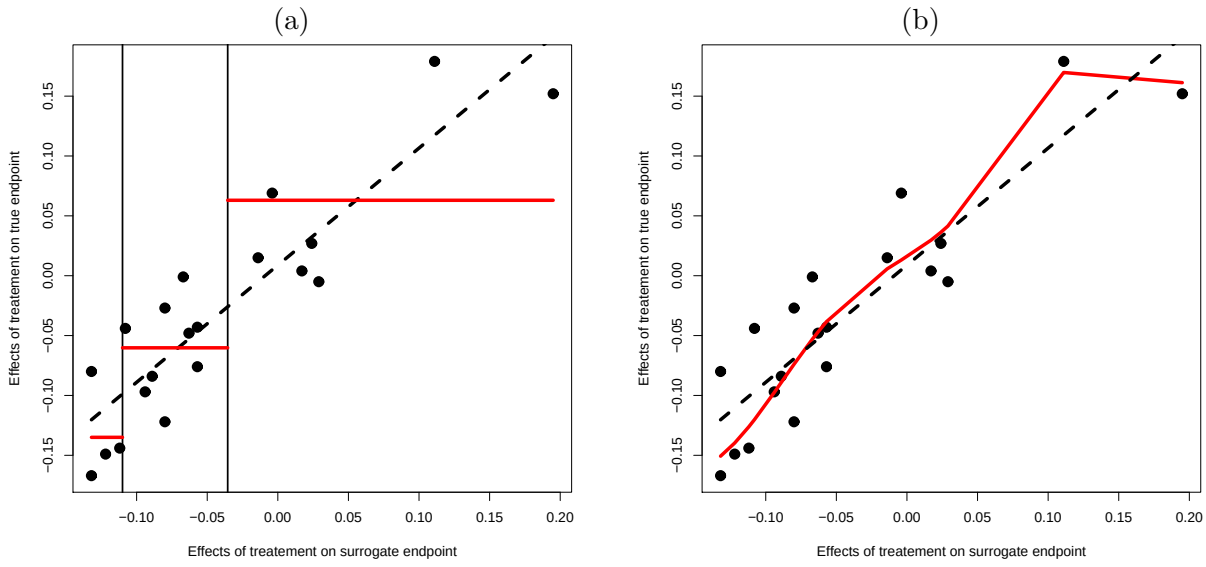


Figure 4: *Clinical Studies in Schizophrenia. (b) Predictions of the final prune tree (solid lines), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate, together with the linear model predictions (dashed line). (c) Predictions from support vector regression (solid line), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate, together with the linear model predictions (dashed lines).*

Turning to support vector machines, the best kernel was, again, the radial kernel. The predictions of the final model are shown in Figure 3(a).

Comparing the lengths of the bootstrap-based confidence intervals, a somewhat different picture emerges than what was seen with cross-validation. Three methods produce intervals of roughly the same length, around 0.640: random forests, regression trees, and support vector machines. The interval for bagged regression trees is quite a bit lower (0.555), with the linear model at first sight the clear "winner" in this case (0.297). However, this is misleading since we have established the association to be of a different, non-linear nature. It should therefore be discarded in this case, thereby motivating, once more, the use of alternative, more flexible methods.

5.3 Clinical Studies in Schizophrenia

The effects of treatment on both endpoints for the schizophrenic dataset were estimated using the methodology described in Section3.4. Figures 4(a)(b) shows the scatterplot of the estimated effects of treatment on both endpoints, for each country involved in the study.

All point estimates are now considerable, between roughly 0.7 and 0.85 in the non-corrected case, and between 0.5 and 0.75 when cross-validation is applied. Figure 4(b) shows the fitted values obtained with regression trees, together with the predictions from the linear model. Also here, the radial kernel did best for the support vector machine method. The final model is shown in Figure 4(c).

We now obtain a very wide interval for support vector machines (0.794), followed by regression trees (0.638). The other three are more narrow, going from 0.455 for random forests, over 0.403 for the linear model, to the winner in this case, bagged regression trees, which produces 0.383.

6 Discussion and Recommendation

In this paper, we have investigated several issues related to the estimation of trial level surrogacy.

First, there is the issue related to the assumption of linear association between the effect of treatment on the true and surrogate endpoints, which can be dealt with by using more flexible modeling techniques that allows other type of association. The methodology developed in this field is not restricted to the linear association between this effects of treatment and here we proposed a more flexible approach that allows to predict the treatment effect on the true endpoint even if the association is not linear.

Second, the use of the delta method to calculate confidence intervals is not recommendable since it makes assumptions valid only in very large samples. Not only are the intervals always symmetric, even when the corresponding distribution is not, it may produce range-violating limits. We therefore considered alternatives: bootstrap methods in general and, for the linear model, Ding's approach. When both of these are applied, they produce similar results.

Third, we have seen that, when no cross-validation based correction is applied, overly optimistic trial-level surrogacy estimates will be found. Many values reported in the recent literature on surrogate marker evaluation ought to be revisited in the light of this observation. Since the differences can indeed be considerable, cross-validation is highly recommendable.

The alternative approaches proposed here generally perform better than the classical ones. For the advanced colorectal cancer studies, we even found the trial-level surrogacy is considerably different from what has been reported in the literature (Burzykowski, Molenberghs and Buyse, 2005). In terms of point

estimates, the alternative approaches typically produce larger values and narrower confidence intervals. This is true, not so much for support vector machines and regression trees, but strongly so for the bagging and random forest methods. These methods in particular are highly recommendable.

Acknowledgments

The authors gratefully acknowledge support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data" and Belgian IUAP/PAI network # P6/03 "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data" of the Belgian Government (Belgian Science Policy).

References

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Alonso, A., Geys, H., Kenward, M.G., Molenberghs, G., and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal* **45**, 1–15.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2004). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology for repeated measurements. *Submitted for publication.*

Alonso, A. and Molenberghs, G. (2006). Surrogate marker evaluation from an information theory perspective. *Submitted for publication.*

Breiman L., Friedman J.H., Olshen R.A., and Stone C.J., 1984. *Classification and regression trees*. New York: Chapman & and Hall/CRC.

Breiman, L. (1996a). Bagging predictors. *Machine Learning* **26**, 123–140.

Breiman, L., (1996b). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350-2383.

Breiman L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Burzykowski, T., Molenberghs, G, and Buyse, M. (2004). The validation of surrogate endpoints by using data from randomized clinical trials: A case study in advanced colorectal cancer. *Journal of the Royal Statistical Society, Series A* **167**, 103–124.

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**, 186–201.

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**, 1–19.

Corfu-A Study Group (1995). Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin for advanced colorectal cancer. *Journal of Clinical Oncology* **13**, 921–928.

Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis* **47**, 537–563.

Daniels, M.J., and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1965–1982.

Ding, C.G. (1996). On the computation of the distribution of the square of the sample multiple correlation coefficient. *Computational Statistics and Data Analysis* **22**, 345–350.

Fletcher R. (1989). *Practical Methods of Optimization.* New York: John Wiley.

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

Gail, M.H., Pfeiffer, R., Van Houwelingen, H.C., and Carroll, R. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.

Galecki, A.T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics* **23**, 3105–3120.

Greco, F.A., Figlin, R., York, M., Einhorn, L., Schilsky, R., Marshall, E.M., *et al.* (1996). Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. *Journal of Clinical Oncology* **14**, 2674–2681.

Ihaka R. and Gentleman R. (1996). R: A Language for data analysis and graphics. *Journal of Computional and Graphical Statistics* **5**, 299–314.

Kay, S.R., Opler, L.A., and Lindenmayer, J.P. (1988). Reliability and validity of the positive and negative syndrome scale of shizophrenics. *Psychiatry Research* **23**, 99–110.

Liaw A. and Wiener M. (2002). Classification and regression by random forest. *The Newsletter of the R Project* **2/3**, 18–22.

Meyer D. (2001). Support vector machines, the interface to libsvm in package e1071. *The Newsletter of the R Project* **1/3**, 23–26.

Moore D.E., Lees B.G., and Davey S.M. (1991). A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Journal of Environmental Management* **15**, 59-71.

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

Smola, A. (1996). Regression estimation with support vector learning machines. *Master thesis*. Technische Universität München, Munich, Germany.

Therneau T.M. and Atkinson E. J. (1997). An introduction to recursive partitioning using the rpart routines. *Technical Report* **61**, Department of Health Science Research, Mayo Clinic, Rochester, New York.

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.

Vapnik V. and Lerner A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control* **24**, 774-780.

Vapnik V. and Chervonenkis A. (1964). A note on one class of perceptrons. *Automation and Remote Control* **25**.

Vapnik V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

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

Verbyla D.L., (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forestry Research* **17**, 1150-1152.

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.