Alternative methods to evaluate trial level surrogacy

Peer-reviewed author version

CORTINAS ABRAHANTES, Jose; SHKEDY, Ziv & MOLENBERGHS, Geert (2008)
Alternative methods to evaluate trial level surrogacy. In: CLINICAL TRIALS, 5(3). p. 194-208.

DOI: 10.1177/1740774508091677
Handle: http://hdl.handle.net/1942/8456

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

Background: 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. Later, Buyse *et al* (2000) proposed a meta-analytic methodology, producing different methods for different settings, which was further studied by Alonso and Molenberghs (2007), in their unifying approach based on information theory.
Purpose: In this paper, we focus our attention on the trial-level surrogacy and propose alternative procedures to evaluate such surrogacy measure, which do not pre-specify the type of association. A promising correction based on cross-validation is investigated. As well as the construction of confidence intervals for this measure.
Methods: In order to avoid making assumption about the type of relationship between the treatment effects and its distribution, a collection of alternative methods, based on regression trees, bagging, 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. We apply the various strategies to data from three clinical studies: in opthalmology, in advanced colorectal cancer, and in schizophrenia.
Results: The results obtained for the three case studies are compared; they indicate that using random forest or bagging models produces larger estimated values for the surrogacy measure, which are in general stabler and the confidence interval narrower than linear regression and support vector regression. For the advanced colorectal cancer studies, we even found the trial-level surrogacy is considerably different from what has been reported.
Limitations: In general the alternative methods are more computationally demanding, and specially the calculation of the confidence intervals, require more computational time that the delta-method counterpart.
Conclusions: First, more flexible modeling techniques can be used, allowing for other type of association. Second, when no cross-validation-based correction is applied, overly optimistic trial-level surrogacy estimates will be found, thus cross-validation is highly recommendable. Third, the use of the delta method to calculate confidence intervals is not recommendable since it makes assumptions valid only in very large samples. It may also produce range-violating limits. We therefore recommend alternatives: bootstrap methods in general. Also, the information-theoretic approach produces comparable results with the bagging and random forest approaches, when cross-validation correction is applied. It is also important to observe that, even for the case in which the linear model might be a good option too, bagging methods perform well too, and their confidence intervals were more narrow.

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

1 Introduction

Prentice (1989)[1] and Freedman, Graubard and Schatzkin (1992)[2] 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)[2] 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)[3] proposed to use instead the *relative effect* (RE), linking the effect of treatment on both endpoints, and a second measure at the individual level, which measures 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)[4–6] 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. Buyse *et al.* (2000)[5] showed that 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)[4] and Gail *et al.* (2000)[6].

While the proposal is elegant, it suffers from several drawbacks. First, separate developments are necessary for different types of endpoints. Buyse *et al.* (2000)[5] considered normally distributed endpoints. An overview of methods for binary, time-to-event, and longitudinal endpoints can be found in Burzykowski, Molenberghs and Buyse (2005)[7]. The main issue is that, especially the individual-level surrogacy, is captured through a disparate range of measures. 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. Second, estimation within a hierarchically formulated model framework can be challenges, for which simplified model strategies had to be developed (Tibaldi *et al.*, 2003)[8], generally based on replacing a hierarchical analysis by a two-stage alternative, where first trials are analyzed 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. In response to these issues, Alonso and Molenberghs (2007)[9] 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 obtained in the first stage of the two stage approach proposed by Buyse *et al.* (2000)[5]. Oosterlinck *et al.* (1997)[10] stated that, for a prognostic factor to be a surrogate endpoint, it is further required that "the effect of treatment on a surrogate endpoint must be 'reasonably likely' to predict clinical benefit." In other words, an endpoint (S) will be a good surrogate for the true endpoint (T) if the results of a trial using endpoint S can be used to make inferences about the results of the trial if T would have been observed and used as endpoint and this with sufficient precision. To demonstrate surrogacy, a high association between the treatment effects on the surrogate and on the true endpoint needs to be established. This is merely the definition of trial-level surrogacy. Establishing a good surrogate could be of benefit in important ways since the follow-up period could be shortened and/or expenses reduced. This is why, given the importance of this measure, we will focus our attention on the trial-level surrogacy.

While some issues have been addressed, important ones remain. In the present article, we address 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)[11] or one of its simplifications outlined in Tibaldi *et al.* (2003)[8]. One of the assumptions made when these approaches are used is that the type of relation is linear, which in general is not necessarily the case, and the model relies also on the normality assumption of the treatment effect of the true endpoint. 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. In order to avoid making assumption about the type of relationship between the treatment effects and its distribution, 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.

Last but not least, it is important to realize that statistics alone will never be able to decide on the fate of

a candidate surrogate. Indeed, the successful adoption of a surrogate endpoint must be simultaneously statistically and clinically convincing. Evidently, therefore, it must be biologically plausible.

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)[7]. Moreover, different type of endpoints are present in each case study, underscoring the generality of our results. Finally, considering this three case studies, we elude the peril of reporting results that are of interest but too specific to a particular situation.

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)[7].

2.1 The Age-Related Macular Degeneration Study (ARMD)

These data arise from a randomized multi-center clinical trial comparing an experimental treatment (interferon-α at 6 million units daily) to a corresponding placebo in the treatment of patients with age-related macular degeneration (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with macular degeneration progressively lose vision. In the trial, the patients' visual acuity was assessed through their ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The raw patient's visual acuity is the total number of letters correctly read. In addition, one often refers to each line The true endpoint is 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

Consider data from 27 randomized multi-center 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)[12–13] 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)[14] and Burzykowski, Molenberghs and Buyse (2005)[7] 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 a four-category tumor response variable, the distribution of the four tumor response categories: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD) (World Health Organization 1979). In accordance with previous analyzes, only centers with at least 3 patients on each treatment arm are considered. The data include 27 trials, with a total sample size of 4010 patients.

2.3 Clinical Studies in Schizophrenia

This is a meta-analysis of five trials in schizophrenic patients (Alonso *et al.*, 2002)[15], which as such is too small a number of trials to apply the meta-analytic methods. Instead, we will used country as a unit of analysis. Note that the choice of unit is an important issue, and should be carefully considered (Cortiñas *et al.*, 2004)[16]. The true endpoint is Clinician's Global Impression (CGI), 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)[17]). 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 remainder of 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)[5] and 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 $\mu_{S_i} = \mu_S + m_{S_i}$, $\mu_{T_i} = \mu_T + m_{T_i}$, $\alpha_i = \alpha + a_i$, and $\beta_i = \beta + b_i$. The authors assumed that the two endpoints are normally distributed. 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, $(m_{S_i}, m_{T_i}, a_i, b_i)^T$, is assumed to be zero-mean normally distributed with variance-covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Sa} & d_{ab} & d_{bb} \end{pmatrix}. \tag{2}$$

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. A 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. Let us subscript all quantities

pertaining to the particular trial under study with 0. It is easy to show (Buyse *et al.*, 2000)[5] that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \quad (3)$$

$$\text{Var}(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (4)$$

Related to prediction equations (3)–(4), a measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$R^2_{\text{trial (f)}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (5)$$

A good surrogate, *at the trial level*, would have (5) close to 1, which will be associated with a surrogate for which 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 (5) then reduces to

$$R^2_{\text{trial (r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \quad (6)$$

This formula is useful also 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 regressing the true endpoint on the surrogate, perhaps using auxiliary information from the intercept (background effect) and properly taking the information into account, ideally through a two-stage approach.

Not all endpoints in all trials are Gaussian. Indeed, while both are Gaussian in the age-related macular degeneration study, the surrogate is ordinal with a time-to-event true endpoint in the advanced colorectal study, and both endpoints are longitudinally measured in the psychiatric study. In Section 3.3, we briefly outline the methodology of Burzykowski, Molenberghs and Buyse (2004)[14] for the ordinal-survival setting. In Section 3.4, the longitudinal methodology is described (Alonso *et al.*, 2003)[18].

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

Burzykowski, Molenberghs and Buyse (2004)[14] extended the methodology proposed by Buyse *et al.*

(2000)[5] to a K-category ordinal surrogate and a failure-time true endpoint. The authors replaced the first stage model by a bivariate copula model for the true ($Y_{T_{ij}}$) and a latent continuous variable ($Y_{\widetilde{S}_{ij}}$) underlying the surrogate endpoint ($Y_{S_{ij}}$). Specifically, to model $Y_{S_{ij}}$ they proposed the proportional odds model:

$$\text{logit}\{P(Y_{S_{ij}} \leq k|Z_{ij})\} = \gamma_{ik} + \alpha_i Z_{ij}. \tag{7}$$

$F_{Y_{\widetilde{S}_{ij}}}(y_S; z)$ is then the marginal cumulative distribution function of $Y_{\widetilde{S}_{ij}}$, given $Z_{ij} = z$. In the same fashion to model the effect of treatment Z_{ij} on the marginal distribution of $Y_{T_{ij}}$ the authors proposed to use the proportional hazard model:

$$\lambda_{ij}(t|Z_{ij}) = \lambda_i(t) \exp(\beta_i Z_{ij}), \tag{8}$$

where β_i are trial-specific effects of treatment Z and $\lambda_i(t)$ is a trial-specific baseline hazard function. The marginal cumulative distribution function of $Y_{T_{ij}}$, following model (8) with $Z_{ij} = z$, is denoted by $F_{Y_{T_{ij}}}(y_T; z)$.

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

$$F_{Y_{\widetilde{S}_{ij}}, Y_{T_{ij}}}(y_T, y_S; z) = C_\theta \left[F_{Y_{\widetilde{S}_{ij}}}(y_S; z), F_{Y_{T_{ij}}}(y_T; z), \theta \right], \tag{9}$$

where $C_\theta[.,.]$ is a distribution function on $[0,1]^2$ with $\theta \varepsilon \Re$, describing the association between $Y_{\widetilde{S}_{ij}}$ and $Y_{T_{ij}}$. An attractive feature of this model is that the marginal models, proportional odds and proportional hazards models, respectively, and the association model can be selected without constraining each other. Using the joint distribution function (9), with proportional hazard model (8) and proportional odds model (7) (or a suitable modification) as marginal models, it is possible to construct the likelihood function for the observed data.

At the first stage Burzykowski, Molenberghs and Buyse (2004)[14] proposed to use the likelihood function to obtain an estimate of θ and estimates of trial-specific treatment effects α_i and β_i on the surrogate and the true endpoint, respectively. At the second stage, the authors proposed to evaluate the trial-level surrogacy by means of 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)[18] extended the methodology proposed by Buyse *et al.* (2000)[5] to the case where both endpoints are longitudinal. This setting poses important challenges in terms of both finding a model that can accommodate such multivariate structures, as well as new measures that allow for the evaluation of 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 ideas in Galecki (1994)[19], Alonso *et al.* (2003)[18] proposed a joint model at the first stage for both responses:

$$\begin{cases} Y_{S_{ijk}} = \mu_{S_i} + \alpha_i Z_{ij} + g_{T_{ij}}(\xi_{ijk}) + \varepsilon_{S_{ij}}, \\ Y_{T_{ijk}} = \mu_{T_i} + \beta_i Z_{ij} + g_{S_{ij}}(\xi_{ijk}) + \varepsilon_{T_{ij}}, \end{cases} \tag{10}$$

where μ_{S_i} and μ_{T_i} are in agreement with (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 (10) does not preclude the more general case. The random vectors associated with the error for both endpoints are assumed to jointly follow a mean-zero multivariate normal distribution with variance-covariance matrix

$$\Sigma_i = \begin{pmatrix} \sigma_{SS_i} & d_{ST_i} \\ d_{ST_i} & d_{TT_i} \end{pmatrix} \otimes R_i, \tag{11}$$

where R_i reflects a general correlation matrix for the repeated measurements. More details can be found in Alonso *et al.* (2003)[18]. If treatment effect is assumed constant over time, then the R^2_{trial} measure proposed by Buyse *et al.* (2000)[5] 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)[5] 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)[8]. This approach produced 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)[14] proposed as a natural candidate to measure individual level surrogacy the parameter θ, since its value modifies the form of the copula function and, consequently, captures the strength of the association between the surrogate and true endpoints. A drawback of this measure is that, for different copula functions, it may assume values from different domains. They proposed the use of some transformations that can be interpreted similarly to a correlation coefficient, irrespective of the copula function. However, it is possible to choose a copula function such that θ has got a meaningful interpretation. In particular they proposed to use the bivariate Plackett copula (Molenberghs and Verbeke, 2005)[20]. Then, θ can be interpreted as the (constant) ratio of the odds for surviving beyond time t given response higher than k to the odds of surviving beyond time t given response at most k. They noted that, as θ involves comparison of survival times of patients classified according to tumor response, it is likely to be length-biased. There are several methods that can be used to correct for length bias and landmark analysis is one of them. Burzykowski, Molenberghs and Buyse (2004)[14] used an approximate solution, which consisted of excluding patients dying before the landmark time and assuming that all recorded responses had occurred before the landmark. Here, we will focus our analysis on a so-called landmark time of 3 months, given that tumor response is usually assessed 3 month after the beginning of chemotherapy. This produced a trial-level surrogacy of $R^2_{\text{trial (r)}} = 0.15$, using Ding (1996)[21] approach, the 95% confidence interval is $[0; 0.41]$. 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 based on random splines (Verbyla *et al.*, 1999; Alonso *et al.*, 2004)[22–23]. 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)[23]. 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 built upon the two-stage approach proposed by Buyse *et al.* (2000)[5], the extension of Burzykowski, Molenberghs and Buyse (2004)[14] for a categorical surrogate and a time-to-event true endpoint, and the method of

Alonso *et al.* (2004)[23] for repeatedly measured endpoints. The three approaches use, at the second stage, the determination coefficient from the regression linking the effects on treatment from both endpoints [$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$]. Three issues require attention and are presented in more detail in the next subsections. First, estimates of trial-level surrogacy may be over-optimistic since the data used to build and evaluate the model are the same. Second, the relation between these effects can, but does not have to be necessarily linear. Third, delta-method confidence intervals can be misleading, as we will show by comparison. This motivates the use of a generally applicable bootstrap alternative.

To evaluate the trial-level surrogacy, we will use the general formulation proposed by Alonso and Molenberghs (2007)[9] based on an information-theoretic measure of association:

$$R_h^2 = \frac{EP(\beta) - EP(\beta \mid \alpha)}{EP(\beta)}, \tag{12}$$

where $\mathrm{EP}(\beta) = \frac{1}{(2\pi e)^n} e^{2h(\beta)}$ denotes the so-called power entropy of β with density function f and h denotes its entropy defined as $h(\beta) = E[-\log f(\beta)]$. In the conditional case, densities are replaced by their conditional counterparts ($h(\beta|\alpha) = E_\alpha[h(\beta|\alpha = \alpha_i)]$). Ample detail can be found in Alonso and Molenberghs (2007)[9]. This measure of association will also be estimated, and for that we will use the method proposed by Kraskov, Stögbauer and Grassberger (2004)[24] to estimate the mutual information ($I(\alpha, \beta) = h(\beta) - h(\beta|\alpha)$), which is based on entropy estimates from k-nearest neighbor distances, and this estimator of the mutual information is data efficient, adaptive, and has minimal bias. Alonso and Molenberghs (2007)[9] also shown that (12) can be rewritten in terms of the mutual information as $1 - e^{-2I(\alpha,\beta)}$, thus we can expect R_h^2 to be data efficient, adaptive, and, of minimal bias as well. Confidence interval will be calculated based on the bootstrap method. Thousand bootstrap samples with replacement will be created and for each of them the $R_{h_b}^2$ will be calculated, with $b = 1, \ldots, 1000$. The 95% confidence interval will be constructed using the 2.5 and 97.5 percentile of the $R_{h_b}^2$, obtained from the 1000 bootstrap samples.

4.1 Cross-validation to Obtain Adequate Estimates of Trial-level Surrogacy Measure

In classification studies in general error rates are calculated using different correction methods based on bootstrap or cross-validation (Efron and Tibshirani, 1997)[25]. Lendasse, Wertz and Verleysen

(2003)[26] extend these results to regression and compare several model selection methods, based on experimental estimates of their generalization errors. In general, they all lead to similar conclusions. Lendasse, Wertz and Verleysen (2003)[26] also state that bootstrap methods are in typically downward biased with small variance, while cross-validation is almost unbiased, but exhibits more variability. In our particular case, given that we are interested in obtaining unbiased estimates of the trial level surrogacy measure, we opt for using a cross-validation method, in particular a k-fold cross-validation method. The k-fold cross-validation procedure consists of splitting the data into k partitions in which $k-1$ partitions are used to build the model and the left-out sample is then used to predict the treatment effect on the true endpoint. This is repeated k times, finally leading to a vector with the prediction of the treatment effect on the true endpoint for all trials, which will be used to estimate the trial-level surrogacy. Precisely, all alternative methods used to estimate trial-level surrogacy will be subjected to a 10-fold cross-validation correction. Numbers of replication other than 10 were used too, leading to comparable results, which is why only the 10-fold cross-validation results are displayed. A similar approach has been used in this context by Baker (2006)[27], in which a trial is left out and considered as a validation trial. His approach estimate the predicted effect of intervention on the true endpoint in a validation trial, using data from each previous trial with surrogate and true endpoints, as well as data on surrogate endpoint only in the trial of interest; he then takes a weighted average of these estimated predicted intervention effects with weights derived from a random-effects model.

4.2 Flexible Models to Estimate Trial-level Surrogacy Measure

It is important to note that, up to now, trial-level surrogacy has been estimated using linear models at the second stage, but the methodology proposed by Buyse *et al.* (2000)[5] is not restricted to this type of association. The general idea is to enable prediction, for a new trial, of the effect of treatment on the true endpoint, with reasonable precision, knowing the effect of treatment on the surrogate. This is why we propose the use of more flexible regression techniques instead, allowing for a more general functional relationship between the effects of treatment, even though generally we expected to observe monotonic type of relationships. Another option would be to use polynomial type of models instead, but still the assumptions of the model must be satisfied to enable valid inferences, which is why flexible non-parametric modeling might be preferable. We will focus on regression trees (Breiman *et al.*, 1984)[28],

bagging algorithms (Breiman, 1996a,b)[29–30], random forest (Breiman, 2001)[31] and support vector regression (Vapnik, 1995)[32], which in general are most useful when a large number of predictors are available, but they can also be used to explore other types of association between two variables, and additionally they do not assume any particular distribution of the response, which can be exploited as well in this type of scenario. Also, it has been shown (Meyer *et al.*, 2002)[33] via a benchmark study in which 9 different regression methods were applied to 12 different datasets and compared by means of mean squared error (MSE), that in general the methods mentioned above perform better than linear models in term of MSE. Other methods can be used as well, but it was not the scope of the paper to use all possible methods. Instead, we have selected some and establish that other regression techniques can be used in this field of surrogate markers.

We used the R statistical computing environment (Ihaka and Gentleman, 1996)[34] and the R packages RPART version 3.1-27 (Therneau and Atkinson, 1997)[35], randomForest version 4.5-15 (Liaw and Wiener, 2002)[36], and the interface libsvm from e1071 version 1.5-12 (Meyer, 2001)[37].

4.2.1 Regression Tree Analysis

The regression tree methodology is a very well-known and widely used technique (Therneau and Atkinson, 1997)[35]. 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)[26,38]. 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 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 (in our particular case we used 10-fold cross-validation) to prune the overfitted tree to an optimal size (Therneau

and Atkinson, 1997)[35]. Following Breiman Breiman *et al.* (1984)[28], we choose as the "right-sized" tree the smallest-sized, i.e., least complex, tree of which the cross-validation costs do not differ appreciably from the minimum cross-validation costs. In particular, these authors proposed a "1-SE rule" for making this selection, i.e., choose as the "right-sized" tree the smallest-sized tree whose cross-validation costs do not exceed the minimum cross-validation costs plus once the standard error of the cross-validation costs for the minimum cross-validation costs tree. It is worth noting that, given that a cross-validation method has to be used to select the "right-sized" tree to obtain a corrected estimate of trial level surrogacy, the cross-validated R-squared obtained from this process can be used as the adequate trial-level surrogacy measure. It is not needed to redo the cross-validation given that it was already applied. 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)[39]. Details about the methods can be found in Therneau and Atkinson (1997)[35].

The trial-level surrogacy measure that will be employed with regression tree analysis takes the ideas proposed by Alonso and Molenberghs (2007)[9] into account and, for this particular model (for the final tree), (12) can be written as

$$RD_{tree} = \frac{D(\beta) - D(\beta \mid \alpha)}{D(\beta)}, \tag{13}$$

where

$$D(\beta) = \sum_{i=1}^{N} (\beta_i - \bar{\beta})^2 \tag{14}$$

is the deviance or total variability of the true-endpoint treatment effects. Furthermore, $D(\beta \mid \alpha)$ denotes the deviance of the final pruned tree when the information of the surrogate outcome treatment effects are accounted for. Assuming that we have m final nodes $(M_1, M_2, \ldots, M_m)$, $D(\beta \mid \alpha)$ can be calculated as:

$$D(\beta \mid \alpha) = \sum_{h=1}^{m} \left(\sum_{\beta_i \in M_h} (\beta_i - \overline{\beta_{M_h}})^2 \right), \tag{15}$$

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

4.2.2 Bagging Regression Trees

Bagging, a contraction of '**b**ootstrap **agg**regat**ing**', is a technique proposed by Breiman (1996a,b)[29–30] 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. Rather, it is a general technique that can be applied in a wide variety of settings to improve predictions. A basic observation is that deterministic learning algorithms tend to overfit. Bagging avoids this by randomizing the input of these learning algorithms in the hope that directions where overfitting occurs for individual predictions cancel out. Whatever overfitting there might be is averaged out when the combining takes place. The explanation as to why bagging is not affected by overtraining is the following. In bagging, the training data set is modified randomly and independently at each step. The bootstrap resampling used in bagging is known as a robust technique. Therefore, in expectation, the distribution of a bootstrap sample, which consists of 63,2% of training data, becomes similar to the real data distribution. In this way, bagging is prevented from overtraining. Details about the method can be found in Breiman (1996a)[29].

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 generated.

4.2.3 Random Forests (RF)

The random forest method (Breiman, 2001)[31] is a supervised learning algorithm that has previously been applied successfully 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 RTA 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 computes proximities between pairs of cases that can be used in clustering, locating outliers, or by scaling, giving useful views of the data; (4) it is well known that random forests avoid overfitting and it has been demonstrated to have excellent performance in comparison to other machine learning algorithms (Breiman, 2001; Svetnik *et al.*, 2003; Meyer *et al.*, 2003)[31,40,41]. Random forests are an effective tool in prediction. Because of the Law of Large Numbers they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Details about random forest can be found in Breiman (2001)[31]. The trial-level measure of surrogacy will be computed similarly to the case in which bagging methods were used.

It is important to mention that we have used cross-validation methods to obtain the trial-level surrogacy measure, to provide a uniform message. However, for this particular case it is not really needed. Since each tree is grown from a bootstrapped sample, on average, about one-third of the observations in the data set will not be used to grow the tree. These observations are considered the out-of-bag (oob) observations for that tree. These oob observations form a natural test set for each tree. Thus, if the oob value for the R-squared is used instead as an estimate for the trial-level surrogacy, it will already correct for overoptimism.

4.2.4 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 throughout a variety of applications. In particular, in regression and time-series prediction applications, excellent performance has been obtained (Drucker *et al.*, 1997; Müller *et al.*, 1997; Stitson *et al.*, 1999; Mattera and Haykin, 1999)[42–45]. 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 Lerner (1963)[46] and Vapnik and Chervonenkis (1964)[47], 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)[48]. The loss function must be modified to include a distance measure. SVM regressions

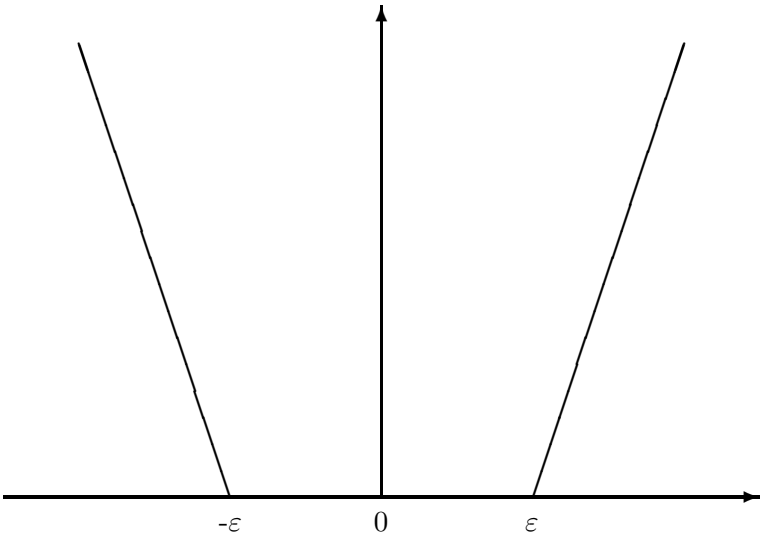

Figure 1: *A piecewise linear ε-insensitive loss function.*

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: $-\varepsilon \leq \omega \cdot \alpha_i - b - \beta_i \geq \varepsilon$. 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:

$$\beta_i - \langle w, \alpha_i \rangle - b \leq \varepsilon, \qquad \langle w, \alpha_i \rangle + b - \beta_i \geq \varepsilon.$$

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:

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

subject to the following constraints:

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

We then need to solve a constrained optimization problem. 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 non-linear functions. Hence, we will use a standard dualization method utilizing Lagrange multipliers, as described in Fletcher (1989)[49]. More details can be found in Vapnik (1995)[32]. Several kernels can be used; we focus on:

- Polynomial: $(\gamma\langle\alpha_i, \alpha_j\rangle + \delta)^d$;
- Radial basic function (RBF): $\exp(-\gamma\|\alpha_i - \alpha_j\|^2)$;
- Sigmoid: $\tanh(\gamma\langle\alpha_i, \alpha_j\rangle + \delta)$.

To select a kernel, all three kernels were tuned, using cross-validation, and, finally, the kernel, together with the set of parameters that produce the smallest mean squared error is retained. In this way, we controlled for the risk of overfitting, given that the set of parameters used to obtain the final model are selected using a cross-validation procedure. We then go on and evaluate the model performance for each of the observations left out in the cross-validated samples and thus the ability of the model to generalize beyond the fitting data.

In this paper, as Hsu *et al.* (2001)[50] pointed out, our choice is the RBF kernel, which can handle the non-linear mapping and has few parameters to be controlled (C between 0.25 and 6, with step of 0.25 and γ between 0.5 and 50 with step of 0.5). The parameters C and γ obtained from the tuning process were then used to estimate the trial-level surrogacy.

Similar to the case of regression trees, 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 on the true endpoint:

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

Here, $D(\beta)$ can be calculated using (14), 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.

4.3 Confidence Intervals for Trial-level Surrogacy Measure

Three types of confidence interval have been used in the past when evaluating trial-level surrogacy. The first one is based on the delta method. The variance of the $R^2_{\text{trial (r)}}$ is computed and the lower and upper bound are calculated as 1.96 times the estimated standard error. The delta method was used first and treated the determination coefficient as a function of the correlation $R_{\text{trial (r)}}$:

$$\text{Var}(R^2_{\text{trial (r)}}) \approx 4R^2_{\text{trial (r)}}\text{Var}(R_{\text{trial (r)}}). \tag{16}$$

Then, using the fact that the variance of Fisher's transformation $Z_F = \frac{1}{2}\ln\left(\frac{1+R_{\text{trial (r)}}}{1-R_{\text{trial (r)}}}\right)$ is approximately equal to $\frac{1}{N-3}$ (Anderson, 1958, p.78)[51] and rewriting $R_{\text{trial (r)}}$ as a function of Z_F

$$R_{\text{trial (r)}} = \frac{e^{2Z_F}-1}{e^{2Z_F}+1},$$

we can now use the delta method and treat $R_{\text{trial (r)}}$ as a function of Z_F. We then get

$$\text{Var}(R_{\text{trial (r)}}) \approx \frac{(1-R^2_{\text{trial (r)}})^2}{N-3}. \tag{17}$$

Combining (16) and (17) leads to

$$\text{Var}(R^2_{\text{trial (r)}}) \approx \frac{4R^2_{\text{trial (r)}}(1-R^2_{\text{trial (r)}})^2}{N-3}.$$

Another method is based on Ding's approach (Ding, 1996)[20], which is used specifically for the linear-regression case. This method is based on reporting the 2.5 and 97.5 quantiles of the cumulative distribution function of $R^2_{\text{trial (r)}}$. Ding (1996)[20] proposed to express the cumulative distribution function in terms of central beta densities, where the terms are evaluated recursively; they present a step-by-step algorithm to compute this distribution function.

The third and last method is based on bootstrap samples of the data. We generate 1000 bootstrap samples with replacement, and calculate for each the estimated value of the trial-level surrogacy. The confidence interval is constructed using the 2.5 and 97.5 percentile from the 1000 bootstrap samples. To compute the bootstrap confidence interval for the cross-validation results, a bootstrap sample was created, then the method applied was combined with the cross-validation correction, through division

of the data into 10 subsets, each time one subset was left out and the predicted value of the treatment effect on the true endpoint obtained for this particular subset. This process was repeated until all subsets are left out and a complete vector with the prediction of the treatment effect on the true endpoint is obtained. Subsequently, the trial-level surrogacy measure is used for each bootstrap sample and finally, based on these results, the 2.5 and 97.5% quantiles from all bootstrap samples were used to construct the cross-validated bootstrap confidence interval.

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, together with an estimation based on the information-theoretic approach using Kraskov, Stögbauer and Grassberger (2004)[24] method to estimate the mutual information. 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 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

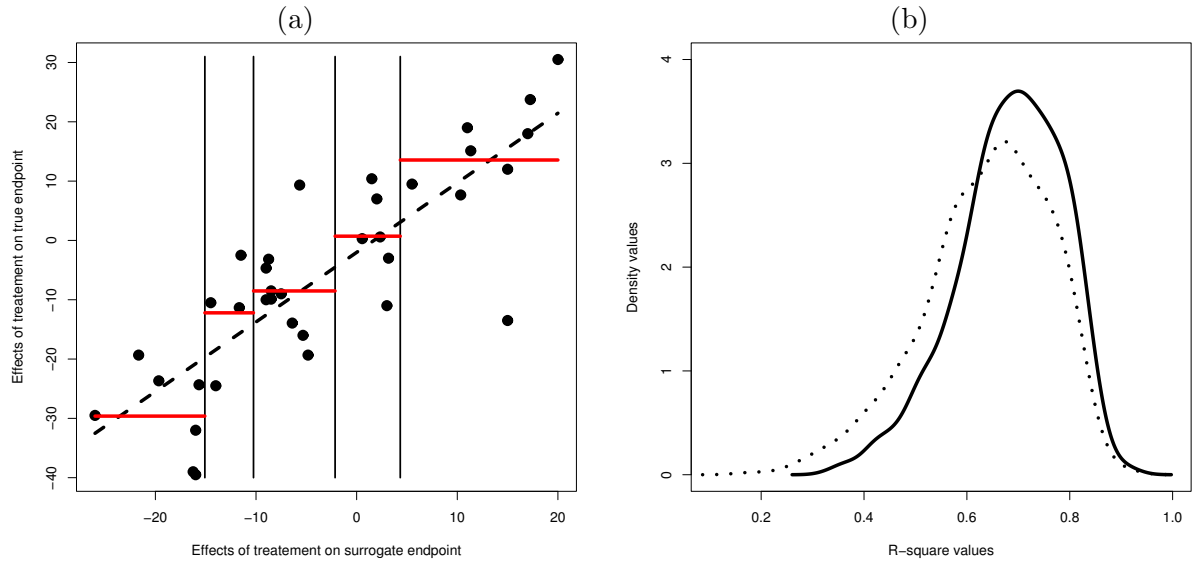

Figure 2: *Age-related macular degeneration study. (a) Scatterplot of the estimated treatment effects for both endpoint in each center, overlaid by predictions of the final pruned tree (solid lines), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the 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).*

forests (0.400). Finally, the shortest interval is found with bagged regression trees (0.347). It is also important to highlight that the lower bound in this particular case is around 0.57, indicating a stronger association, compared to the linear-model counterpart.

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 the random-forest method is used.

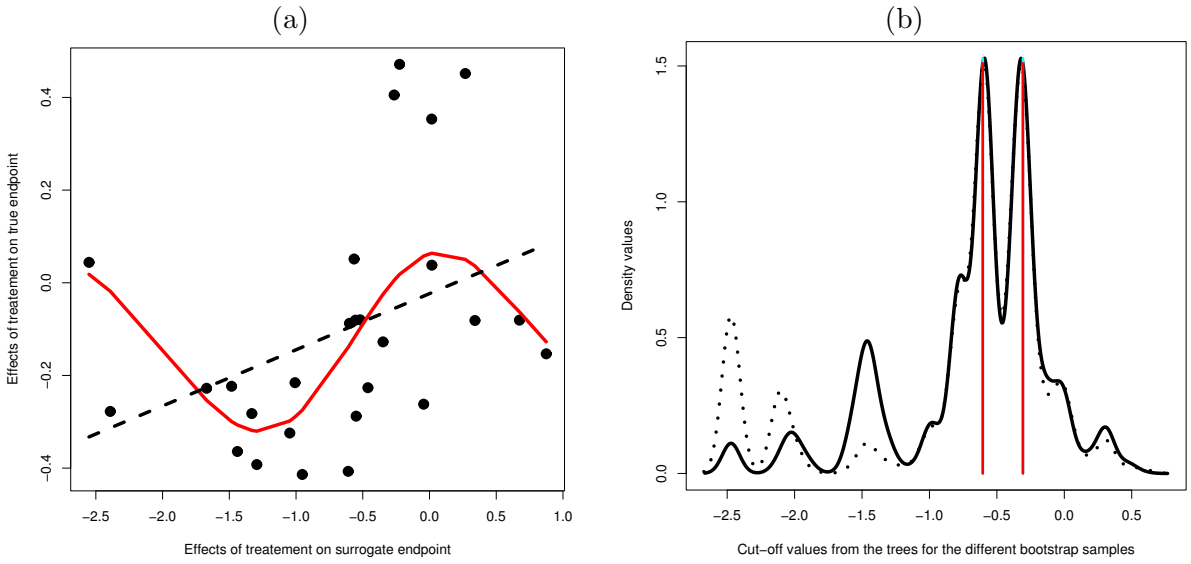

Figure 3: *Advanced Colorectal Cancer. (a) Scatterplot of the estimated treatment effects for both endpoint in each center, overlaid by predictions from support vector regression (solid line), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate, 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.*

Here, we are referring to those values (cut-off) used to split each of the nodes for each of the 1000 trees constructed. 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.

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 can be misleading since we have established the possibility of association to be of a different, non-linear nature, based on the results obtained from the lack-of-fit test ($F_{8,17} = 17.92$, $p < 0.0001$). It should therefore be discarded in this case, thereby motivating, once more, the use of

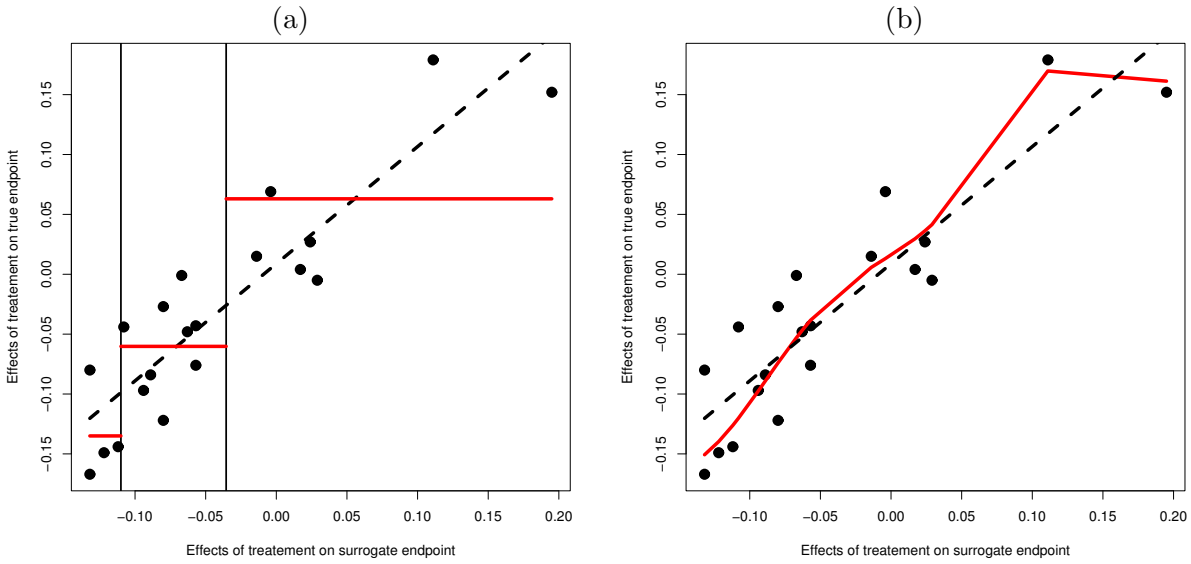

Figure 4: *Clinical Studies in Schizophrenia. (b) Predictions of the final prune tree (solid lines), modeling the effect of the treatment on the true endpoint against the effect of the treatment on the surrogate, together with the linear model predictions (dashed line). (c) Predictions from support vector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

alternative, more flexible methods. This is in line with the observation made from Figure 3 that the impact on the true, survival outcome depends on the range of the surrogate. It is important to recall that the four-point surrogate is an ordinal indication for tumor response. Through dichotomizing the ordinal outcome in the three possible ways, Burzykowski, Molenberghs and Buyse (2004)[14] had found that the four-point outcome exhibits non-trivial behavior. The dichotomy 12/34 turned out to be more informative than the original outcome, with the reverse happening for the 1/234 and 123/4 dichotomous. This at first sight aberrant behavior is arguably caused by the fact that the 12/34 dichotomy essentially is between growth and shrinkage of the tumor, which is relatively easy to establish, whereas the other two are more subtle distinctions in terms of rates of growth and shrinkage. Arguably, these issues are a recipe for non-linear relationships with the true endpoint. Of course, we recognize this may neither be the best nor the entire explanation for such non-linear behavior. The key point, though, is that the methodology is able to cope with such less than straightforward behavior.

5.3 Clinical Studies in Schizophrenia

The effects of treatment on both endpoints for the schizophrenic dataset were estimated using the methodology described in Section 3.4. Figures 4(ab) 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.

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 Recommendations

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, allowing for other type of association. The methodology developed in this field is not restricted to the linear association between this type of effects of treatment and, in response to this, we proposed a more flexible approach, enabling predicting of the treatment effect on the true endpoint even if the association is non-linear. From a practical, clinical standpoint, it might be hard to conceive of non-linear relationships that are non-monotone and, therefore, even when linearity would not hold. Fortunately, the elegance of the proposed methodology is its flexibility with which the data can be modeled and therefore used to discern the precise nature of the relationship. We believe it is important to not rule out non-monotone relationship a priori since, when they would occur, it is imperative for the researchers to scrutinize such a phenomenon.

An important consequence of allowing for non-linear relationships is that, owing to the vast number of possible functional forms that come within reach, often relatively parsimonious yet flexible parametric

shapes can be employed. For example, relatively simple power models might sometimes, but of course not always, outperform conventional linear predictors that are of a polynomial structure in the covariates.

Second, 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.

Third, 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.

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)[7]. 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. Also, the information-theoretic approach produces comparable results with the bagging and random forest approaches, when cross-validation correction is applied.

It is also important to observe that, even for the case in which the linear model might be a good option too, as in two of the case studies presented here, bagging methods perform well too, and their confidence intervals were more narrow.

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

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

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

3. Buyse, M., and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 186–201.

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

5. 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.

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

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

8. 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.

9. Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics* **63**, 180–186.

10. Oosterlinck W., Mattelaer J., Casselman J., Van Velthoven R., Derde M.P., Kaufman L. (1997) PSA evolution: a prognostic factor during treatment of advanced prostatic carcinoma with total androgen blockade. Data from a Belgian multicentric study of 546 patients. *Acta Urol Belg* **65**,63–71.

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

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

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

21. 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.

22. 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.

23. 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.*

24. Kraskov, A., Stögbauer, H., and Grassberger P. (2004). Estimating Mutual Information. *Physical Review E* **69**, 066138.

25. Efron B. and Tibshirani R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association***92**, 548–560.

26. Lendasse, A., Wertz, V., and Verleysen, M. (2003) ICANN 2003, Joint International Conference on Artificial Neural Networks, June 26-29, 2003, Istanbul (Turkey). Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003, O. Kaynak, E. Alpaydin, E. Oja, L. Xu eds, Springer-Verlag, Lecture Notes in Computer Science 2714, 2003, pp. 573–580.

27. Baker S.G., 2006. A simple meta-analytic approach for binary surrogate and true endpoints. *Biostatistics*, **7**, 57–70.

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

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

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

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

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

33. Meyer, D., Leisch, F., Hornik, K., 2002. Benchmarking support vector machines. *Technical Report* bfseries 78, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

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

35. 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.

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

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

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

39. Svetnik,V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inform. Comput. Sci.*,**43**, 1947–1958.

40. 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.

41. Meyer, D., Leisch, F., Hornik, K., 2003. The support vector machine under test. *Neurocomputing*,**55**, 169–186.

42. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik V. 1997. Support vector regression machines. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, pp. 155–161.

43. Müller, K.R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik V. 1997. Predicting time series with support vector machines. In: Gerstner W., Germond A., Hasler M., and Nicoud J.-D. (Eds.), Artificial Neural Networks ICANN'97, Berlin. Springer Lecture Notes in Computer Science **1327**, 999–1004.

44. Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. 1999. Support vector regression with ANOVA decomposition kernels. In: Schölkopf B., Burges C.J.C., and Smola A.J. (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press Cambridge, MA, pp. 285-292.

45. Mattera, D. and Haykin, S. 1999. Support vector machines for dynamic reconstruction of a chaotic system. In: SchölkopfB., Burges C.J.C., and Smola A.J. (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, MA, pp. 211–242.

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

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

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

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

50. Hsu C.W., Chang C.C., Lin C.J. (2001). A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

51. Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

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 (splitting the dataset into 10 subset of similar size, ARMD (6 subsets with 4 and 4 with 3 observations), colorectal (7 subsets with 3 and 3 with 2 observations) and schizophrenia (with ten subsets of 2 observations)). 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
Information Theory Approach	0.726 [0.567;0.875]	0.298 [0.109;0.673]	0.641[0.355;0.791]
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]