Information theory-based surrogate marker evaluation from several randomized clinical trials with binary endpoints, using SAS

Peer-reviewed author version

TILAHUN ESHETE, Abel; ASSAM NKOUIBERT, Pryseley; ALONSO ABAD, Ariel & MOLENBERGHS, Geert (2008) Information theory-based surrogate marker evaluation from several randomized clinical trials with binary endpoints, using SAS. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 18(2). p. 326-341.

DOI: 10.1080/10543400701697190
Handle: http://hdl.handle.net/1942/9581

Information-theory Based Surrogate Marker Evaluation from Several Randomized Clinical Trials with Binary Endpoints, Using SAS

Abel Tilahun **Assam Pryseley** **Ariel Alonso**

Geert Molenberghs

Hasselt University, Center for Statistics, Agoralaan 1, 3590 Diepenbeek, Belgium

Abstract

One of the paradigms for surrogate marker evaluation in clinical trials is based on employing data from several clinical trials: the meta-analytic approach. Originally developed for continuous outcomes by means of the linear mixed model, also other situations are of interest. One such situation is when both outcomes are binary. While joint models have been proposed for this setting, they are cumbersome in the sense of computationally complex and of producing validation measures that are, unlike in the Gaussian case, not of an R^2 type (Burzykowski, Molenberghs, and Buyse 2005). A way to put these problems to rest is by employing information theory, already applied in the continuous case (Alonso and Molenberghs 2007). In this paper, the information-theoretic approach is applied to the case of binary surrogate and true endpoints. Its use is illustrated using a case study in acute migraine and its performance, relative to existing methods, assessed by means of a simulation study. Since the usefulness of a method critically depends, among others, on the availability of software, a SAS implementation accompanies the methodological work.

Some Key Words: Hierarchical model; Meta-analysis; Pseudo-likelihood; Random-effects model; Surrogate endpoint

1 Introduction

The use of surrogate endpoints in clinical trials is increasing, necessitating the development of sound statistical methods for use in the evaluation process. The following definitions are in common use (Biomarkers Definition Working Group 2001). A *clinical endpoint* is a characteristic or variable that reflects how a patient feels, functions, or survives. A *biomarker* is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. A *surrogate endpoint* is a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit, harm, or lack thereof.

One important reason for the present interest in surrogate endpoints is the advent of a large number of biomarkers that closely reflect the disease process. An increasing number of new drugs have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Ferentz 2002). There is increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint (Lesko and Atkinson 2001).If the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now. It is therefore best to use *validated* surrogates, though one needs to reflect on the precise meaning and extent of validation (Schatzkin and Gail 2002). Like in many clinical decisions, statistical arguments will play a major role, but ought to be considered in conjunction with clinical and biological evidence. At the same time, surrogate endpoints can play different roles in different phases of drug development. While it may be more acceptable to use surrogates in early phases of research, one should be much more restraint using them as substitutes for the true endpoint in pivotal phase III trials, since the latter might imply replacing the true endpoint by a surrogate for all future studies as well, a far-reaching decision. For a biomarker to be used as a "valid" surrogate, a number of conditions must be fulfilled. The ICH Guidelines on Statistical Principles for Clinical Trials state that "In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome" (International Conference on Harmonisation 1998).

While initially done in the context of a single trial (Prentice 1989, Freedman, Graubard, and Schatzkin 1992, Buyse and Molenberghs 1998), the meta-analytic framework is now a well accepted one. It allows to cast the evaluation in terms of two important concepts and ultimately quantities: trial-level and individual-level surrogacy. Several authors have contributed to its development; a synthesis is provided in Burzykowski, Molenberghs, and Buyse (2005). Several issues still surround the framework. First, while reasonably feasible for continuous, normally distributed endpoints, thanks to the availability of the

linear mixed model (Verbeke and Molenberghs 2000), the non-Gaussian case is less straightforward, be it for surrogate and true outcomes from the same type or of a different nature. And then even the Gaussian case already needed special attention (Tibaldi *et al* 2003). The main issue is that joint models for the outcomes are needed, preferably allowing for the hierarchy induced by disposing of many trials. Second, and related to the previous comment, is that different models produce different validation quantities. While for the Gaussian case one is led to R^2 type measures, for other settings such measures as odds ratios and Kendall's τ crop up, with the situation getting more complex for longitudinally measured endpoints.

Alonso and Molenberghs (2007), in an effort to alleviate the reported issues, adopted an information theory approach (ITA) to propose an evaluation method that is simple in both use and interpretation of the resulting measures. They applied their ideas to longitudinally measured, continuous endpoints. In this paper, we will adapt the framework to the situation where both outcomes are binary. The methodology will be introduced in Sections 3 as far as general concepts are concerned and in Section 4 specifically for ITA, and then put to the test in a simulation study (Section 5) and exemplified by means of a case study in acute migraine, introduced in Section 2 and analyzed in Section 6. The work reported here is supplemented with a generic implementation as a SAS macro, available to the interested reader from the authors,

2 A Meta-analysis of Ten Clinical Trials in Acute Migraine

Consider a meta-analysis of 10 early phase trials assessing the efficacy of several therapies for the treatment of acute migraine crises. Each trial was placebo-controlled and aimed at evaluating one of three experimental treatments. Two trials also included an active control arm. Overall, 801 patients were available, recruited over 38 different centers, with between 1 and 86 patients enrolled per center. Severity of headache and migraine-related symptoms were measured prior to and at several occasions after the dose administration. Severity was rated on a four-grade intensity scale (0 =no, 1 =mild, 2 =moderate, 3 =severe). Clinically relevant endpoints for efficacy included pain-free (pain score=0) and pain relief (pain score$<= 1$) two hours post-dose. The main goal is to identify what symptoms typically associate with migraine episodes, such as, for example, nausea, vomiting, increased sensitivity to light, i.e., photophobia, as well as to sound, i.e., phonophobia.

3 Validation Methods

We review the meta-analytic approach for two binary variables. In line with Renard *et al* (2002), it is assumed the binary variables result from dichotomized normally distributed ones, thence concepts of the continuous meta-analytic framework (Buyse *et al* (2000) can be employed. Extensions of the framework to other settings are brought together in Burzykowski, Molenberghs, and Buyse (2005). Motivated by the computational complexity of this framework, we will first discuss a number of simplifying strategies (Tibaldi *et al* 2003) and then move on to the information-theoretic approach of Alonso and Molenberghs (2007).

3.1 The Meta-Analytic Approach for Binary Endpoints

To extend the methodology used for continuous endpoints to the case of binary endpoints , Renard *et al* (2002) adopted a latent variable approach, resting on the assumption that the observed binary variables result from dichotomizing an unobserved continuous variable based on the threshold chosen. Assume a pair of latent variables $(\widetilde{S}_{ij}, \widetilde{T}_{ij})$, representing the continuous, underlying values of the surrogate and true endpoints for subject $j = 1, \ldots, n_i$ in trial $i = 1, \ldots, N$, following a random-effects model at the latent scale:

$$\widetilde{S}_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \quad (1)$$

$$\widetilde{T}_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}, \quad (2)$$

μ_S and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random (i.e., trial-specific) intercepts, a_i and b_i are random treatment effects, and ε_{Sij} and ε_{Tij} are error terms. The random effects are zero-mean normally distributed with covariance matrix

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

The error terms are also zero-mean normally distributed with covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{ST} \\ \rho_{ST} & 1 \end{pmatrix}.$$

The implied model for the observed binary outcomes is

$$\Phi^{-1}[P(S_{ij}=1|m_{Si},m_{Ti},a_i,b_i)] = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij}, \quad (3)$$

$$\Phi^{-1}[P(T_{ij}=1|m_{Si},m_{Ti},a_i,b_i)] = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij}, \quad (4)$$

where Φ denotes the standard normal cumulative distribution function. Formulation (1)–(3) allows the use of the coefficient of determination

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

as the trial-level R^2, whereas the individual-level R^2_{indiv} is equal to the square of ρ_{ST}. These quantities are unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, belong to the unit interval. A surrogate could thus be adopted when R^2_{trial} is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and biopharmaceutical judgment involved in the decision process. It is important to be aware that, rather than solely relying on point estimates, also confidence intervals ought to play a role in such cutoffs. Note that, here, trial is considered as experimental unit which can be replaced by center, investigator or any other suitable experimental unit, depending on the nature of the study conducted. The issue of the unit of analysis has been thoroughly studied by Cortiñas *et al* (2004) and Tilahun *et al* (2007).

After fitting such models, we thus obtain two measures of surrogacy, captures as trial-level and individual-level coefficients of determination. A surrogate could be adopted, for clinical trial purposes, when R^2_{trial} is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and other judgment involved in the decision process. The $R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}}$ is based on $\mathbf{\Sigma}$.

3.2 Parameter Estimation

Model (1)–(3) belongs to the class of so-called generalized linear mixed models (Molenberghs and Verbeke 2005), even though the logit link is more generally used for binary outcomes. Molenberghs and Verbeke (2005) discuss a variety of commonly used estimation methods, including maximum likelihood with numerical integration over the random effects, penalized quasi-likelihood, marginal pseudo-likelihood,

and Laplace approximation. These methods suffer to various extents from computational complexity and severe bias (Rodríguez and Goldman 1995, Molenberghs and Molenberghs 2005). For the specific case of the probit link, as in (3)–(4), Renard *et al* (2002) have suggested the use of so-called maximum pairwise likelihood (MPL), a form of pseudo-likelihood (Molenberghs and Verbeke 2005). Let us describe this method.

Assembling all parameters into the vector $\boldsymbol{\Theta}$, the contribution of the ith trial ($i = 1, \ldots, N$) to the likelihood, conditional on $\boldsymbol{b_i} = (m_{Si}, m_{Ti}, a_i, b_i)^T$, is

$$L_i(\boldsymbol{\Theta}|\boldsymbol{b_i}) = \prod_{j=1}^{n_i} P(S_{ij}, T_{ij}|\boldsymbol{b_i}). \tag{6}$$

Maximum likelihood estimation follows from integrating (6) over $\boldsymbol{b_i}$, summing over all subjects, taking the logarithm, and maximizing

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \ln \int L_i(\Theta|\boldsymbol{b_i})\phi(\boldsymbol{b_i}; \boldsymbol{D})d\boldsymbol{b_i} \tag{7}$$

over $\boldsymbol{\Theta}$. Here, $\phi(\boldsymbol{b_i}; \boldsymbol{D})$ denotes the mean-zero multivariate normal density with covariance matrix $\boldsymbol{D}$. The intractable nature of (7) dictates the use of one or other form of approximation, as mentioned earlier. Renard *et al* (2002) suggested the use of maximum pairwise likelihood (MPL), a pseudo-likelihood approach based on replacing the likelihood by a product of conditional and/or marginal densities. In our particular case, the proper likelihood contribution of trial i is replaced by all possible pairwise margins. Detailed overviews of the methodology can be found in Molenberghs and Verbeke (2005) and Burzykowski, Molenberghs, and Buyse (2005).

3.3 Drawbacks and Simplified Modeling Strategies

While it is technically possible to fit the bivariate probit model, the use of which necessitated by the pairwise likelihood approach, there still are a number of drawbacks associated with the approach outlined. First, the resulting surrogate marker evaluation measures apply to the postulated latent variables rather than to the observed binary variables. Second, the computational burden still is considerable. Third, the approach might result in an ill-conditioned variance-covariance matrix, thence calling the reliability of the association measures derived into question. Some of these problems occur in the continuous case as well, for a discussion of which we refer to Tilahun *et al* (2007). To address the problem of

computational burden, Tibaldi *et al* (2003) suggested several simplifications for the case of continuous true and surrogate endpoints. They have organized their simplifications along three so-called *dimensions*. Let us outline these in turn and then examine their usefulness for the binary case.

The *trial dimension* is concerned with whether the random effects are considered fixed or random. If the trial-specific effects are chosen to be fixed, a two-stage approach is effectively adopted. The first-stage model will take the form

$$g_S[P(S_{ij}=1)] = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{8}$$

$$g_T[P(T_{ij}=1)] = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{9}$$

where g_S and g_T are appropriate link functions. At the second stage, the estimated treatment effect on the true endpoint is regressed on the treatment effect and intercept from the surrogate endpoint model:

$$\widehat{\beta}_i = \widehat{\lambda}_0 + \widehat{\lambda}_1 \widehat{\mu}_{Si} + \widehat{\lambda}_2 \widehat{\alpha}_i + \varepsilon_i. \tag{10}$$

Here, $\widehat{\mu}_{Si}$, $\widehat{\alpha}_i$, and $\widehat{\beta}_i$ are estimates obtained from (8) and (9), with the λ being regression coefficients estimated using conventional linear regression techniques. The trial-level $R^2_{\text{trial(f)}}$ then is the coefficient of determination obtained by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{Si}$ and $\widehat{\alpha}_i$, whereas $R^2_{\text{trial(r)}}$ is obtained from the coefficient of determination resulting from regressing $\widehat{\beta}_i$ on $\widehat{\alpha}_i$ only. The subscripts (f) and (r) refer to 'full' and 'reduced', respectively. The individual-level measure of surrogacy is then calculated using ITA, to be outlined in Section 4.

The second option is to consider the trial-specific effects as random. How one then proceeds is related to the so-called *endpoint dimension*. Indeed, though natural to assume the two endpoints correlated, this choice does increase computational complexity. The desirability to accommodate the bivariate nature of the outcome is associated with interest in R^2_{indiv}, which is in some cases of secondary importance. This applies, for example, when the trial-level surrogacy R^2_{trial} is the sole quantity of importance to the investigator; this may occur in a clinical trial context, as opposed to when the surrogate is used, for example, for diagnostic purposes. Furthermore, even when the individual-level surrogacy is of importance, there now is the possibility to estimate it by making use of ITA rather than by formulating a bivariate model.

If in the trial dimension, the trial-specific effects are considered fixed, then models (8)–(9) are fitted

separately. Similarly, if the trial-specific effects are considered random, then the corresponding models

$$\begin{aligned} g_S[P(S_{ij}=1)] &= \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij}, \\ g_T[P(T_{ij}=1)] &= \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} \end{aligned}$$

are fitted separately, i.e., the corresponding error terms in the two models are assumed to be independent. Otherwise, the outcomes are considered correlated and a full mixed-modeling approach is followed.

Except when a bivariate mixed-modeling approach is followed, there is a need to adjust for the heterogeneity in the amount of information contributed by the various trials. This is the subject of the *measurement error dimension*. One can either ignore this phenomenon of weight the trial-specific contributions according to trial size. This then turns (10) into a weighted regression. Tibaldi *et al* (2003) also proposed another measure than just trial size as candidate weights, but we will leave these out of consideration.

4 The Information Theory Approach

As mentioned in the previous section, the main rationale for explicitly accommodating for the two endpoints' correlation is to allow for individual-level surrogacy estimation, at the price of increased computation complexity. It is therefore advantageous to switch towards ITA, which is elegant and computationally simple (Alonso and Molenberghs 2007). Let us describe the method for the case of generalized linear models, obviously containing bivariate outcomes as a particular instance. Consider the following generalized linear models:

$$g_T[E(T_{ij})] = \mu_{Ti} + \beta Z_{ij}, \tag{11}$$

$$g_T[E(T_{ij}|S_{ij})] = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \tag{12}$$

Model (11) relates the expected value of the true endpoint to the treatment only while (12) relates it to surrogate endpoint as well. Upon fitting (11)–(12), the individual-level association can be measured by:

$$R_h^2 = 1 - \frac{1}{N} \sum_{i=1}^{N} \exp\left(\frac{-G_i^2}{n_i} \right), \tag{13}$$

where G_i^2 denotes the likelihood ratio statistics to compare (11) and (12) within trial i, the size of which is n_i. The subscript h is added to distinguish this quantity from earlier uses of R^2 measures. One issue

arising is that, for discrete random variables R_h^2 has an upper bound smaller than one, as shown by Alonso and Molenberghs (2007), who therefore suggested the use of an adjusted version:

$$R^2_{\text{adj}} = \frac{R_h^2}{1 - \exp[-2H(Y)]}, \tag{14}$$

where $H(y)$ is the log-likelihood of the true endpoint divided by the total number of subjects. To augment R^2_{adj} with a measure for uncertainty, Alonso and Molenberghs (2007) suggested asymptotic as well as bootstrap-based intervals.

ITA ideas can be applied to compute the trial-level R^2_{trial} too, using the fully hierarchical model for continuous outcomes (Buyse *et al* 2000). The resulting $R^2_{h,\text{trial}}$ will take the same form as (13), with now G_i^2 the likelihood ratio statistics for comparing models relating treatment effect to the true endpoint, with and without adjusting for the surrogate endpoint. Since this second-stage model is for continuous endpoints, the issue of an upper bound smaller than one does not crop up.

5 Simulation Study

We will now assess the performance of the proposed approach, first laying out the design of our simulation study and then summarizing the results.

5.1 Design of Simulation Study

The data were generated based on model (3)–(4). The parameters were set equal to $\mu_S = 0.5$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. Values assumed for the covariance matrices are:

$$\Sigma = \begin{pmatrix} 3 & 2.4 \\ & 3 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ & 3 & 0 & 0 \\ & & 3 & 2.84605 \\ & & & 3 \end{pmatrix}.$$

After generating continuous outcomes based on the above models, the corresponding binary variables are obtained by dichotomizing the resulting continuous outcomes using the fixed intercepts as cut-off points, setting values exceeding the intercept to 1 and 0 otherwise. These model choices imply R^2_{trial}=0.90 and $R^2_{\text{indiv}} = 0.64$, at the continuous scale. Given we want to compare a variety of estimation methods, it would be unwieldy to expand the R^2 values into grids. However, the performance relative to these value

is of interest, since they are in the range of what is observed for the acute migraine data, as will be shown in Section 6. The important issue can then be raised as to what values should be reached for good surrogacy. These are very difficult to answer purely in statistical terms. However, quantification of surrogacy provides an important piece of evidence when deciding on the adoption of a surrogate; such a decision will clearly also involve biological, biomedical, ethical, and economic arguments.

5.2 Simulation Results

The number of trials was fixed to either 5, 10, 20, or 30. There were 2 sets of trial sizes used, the first set consisting of 10, 20, 40 or 60, which we term *small trial size*. The second set consists of 100, 150, 200 or 300, termed *large trial size*. A full combination of the number of trials and trial sizes was obtained. In each case, 100 runs were performed. We further distinguish between the bivariate and univariate models on the one hand, and mixed- versus fixed-effects models on the other hand. The mixed models take the form of the probit model in the bivariate situation and the GLMM in the univariate case. The simulation results, for a small selection set of numbers of trials and trial sizes, are displayed in Tables 1 and 2. The results shown are representative for those not shown as well.

Let us first consider association at the trial level. The simulation reveal that the full bivariate random effects model and its univariate counterpart are consistent in that both models produce surrogacy measures approaching the true values with the number of trials and the number of subjects increasing. However, the corresponding fixed-effects models lead to underestimation, even for larger sample sizes. Even though the full bivariate model leads to measures at the latent scale, since it measures the association between the treatment effects on the two endpoints, we expect it to be preserved at the explicit scale. This claim is corroborated by the results from the univariate mixed model, which operates at the observed binary scale. It is also noteworthy that there is not much difference between the ITA and the conventional approach of regressing the treatment effect on the true endpoint on the treatment effect on the surrogate endpoint.

Turning to the individual-level association, the full bivariate random-effects and bivariate fixed-effect models result in individual-level measures close to the true value, i.e., the theoretical value at the latent scale. However, they are hard to translate from the latent scale to the explicit one. ITA is

a convenient way out of this problem. An important observation made from the simulation study, that will be confirmed by the data analysis in the next section, is that the values reported with ITA are substantially smaller than their latent counterparts, in line with expectation: switching from the latent scale to the explicitly observed scale reduces association. This is a manifestation of the fact that important information is lost when switching from a continuous to a binary scale. Of course, in a real study, the binary variables are the only ones observed and it is therefore fair to assert that the ITA is a fair representation of reality, whereas the other methods are overly optimistic.

6 Analysis of Acute Migraine Data

Let us analyze the data described in Section 2, using the methods introduced in Sections 3 and 4. Of the symptoms studied: nausea, vomiting, photophobia, phonophobia, the photophobia symptom had the highest trial-level surrogacy. Results for both the trial- and individual-level surrogacy are presented in Table 3. For illustration, we also provide the results for phonophobia (Table 4), which is not retained since the validation measures are consistenly lower. Indeed, while it might be considered acceptable, *if it were the only candidate available*, we now conclude, by comparison, that photophobia is the winner.

Both point estimates as well as 95% confidence intervals are presented. Observe that the univariate and bivariate fixed-effects models result in smaller R^2_{trial} than the random-effects counterpart. However, the latter is unreliable since based on an ill-conditioned covariance matrix, in the sense of a grossly inflated leading eigenvalue. Basing our conclusions on the univariate mixed effect model, in the simulations found to work well, it is fair to assert that, at the trial level, the presence of photophobia is a good surrogate for migraine severity, i.e., the corresponding R^2_{trial} may be considered sufficiently high. The reasonably good agreement between the treatment effects at both levels, and in addition the absence of obvious outliers, is clear from Figure 1, a so-called bubble plot, displaying a scatter of the pairs of treatment effects for each unit. The size of the circles, or bubbles, is proportional to the number of patients per unit.

The R^2_{indiv} for the bivariate fixed and mixed models are higher than their univariate counterparts. This is expected for the same reason as explained in Section 5, i.e., one is at the latent scale, whereas the ITA works at the interpretationally more relevant explicitly observed scale.

7 Discussion

In this paper, we reviewed the meta-analytic strategy by Buyse *et al* (2000), its extension to binary endpoints, and the information theoretic approach for validating surrogate endpoints. The application of the latter framework with binary data is novel. The meta-analytic approach and its extension are mathematically appealing, but encounter practical and/or computational issues. The information theoretic approach involves substantial mathematics yet it is more practically feasible than the meta-analytic approach as it depends on simple univariate models. In addition, the initial meta-analytic framework, where individual-level surrogacy is expressed at the latent probit level, leads to overestimation of the said quantity. Since the ITA operates at the explicitly observed scale, it provides a fairer and more useful quantity.

Additionally, the computational complexity of the full random-effects meta-analytic framework has led to the use of simplifying frameworks, trading the random effects for fixed effects on the one hand and/or bivariate, joint modeling of both endpoints by two univariate, separate models. These simplifications work well when the number of trials and the number of subjects per trials is large, indicating one should in practice carefully consider the unit of analysis. Of course, the choice of unit has also implications in terms of substantive interpretation, as elucidated in Cortiñas *et al* (2004).

Applying the proposed methodology to acute migraine trial data has shown that photophobia is a reasonably good surrogate at the trial level, whereas it surrogacy at the individual level may be called into question. This finding is of interest and may spark of further investigation from a clinical and biopharmaceutical perspective.

Absence of standard software has been one of the limiting factors hampering the use of the meta-analytic approach. The authors have developed a SAS macro for the calculations laid out in the paper, and details are to be found in the Appendix.

Clearly, the use of validation methods, such as the ones proposed in this paper, whether based on R^2, other association measures, or ITA, is but one component of the broader surrogacy picture. All methods discussed here fall within the meta-analytic framework. This is also true for the recent work by Baker (2006), who proposed the use of average prediction error based on relatively straightforward regression

models. Also, regardless of the framework within which one is working, once a surrogate has been adopted, or even before, it is important to assess how it will perform in a new trial. Burzykowski and Buyse (2006) proposed the so-called *surrogate threshold effect* (STE). Their method is intended to derive a sample size large enough for a treatment effect on the surrogate endpoints to translate into a meaningful and significant effect on the true endpoint.

Acknowledgment

The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

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

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

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

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

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

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

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

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.

Ferentz, A.E. (2002). Integrating pharmacogenomics into drug development. *Pharmacogenomics*, **3**, 453–467.

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

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998). ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. *Federal Register*, **63**, No. 179, 49583.

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

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

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

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal*, **44**, 921–935.

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73–89.

Schatzkin, A. and Gail, M. (2002). The promise and peril of surrogate end points in cancer research. *Nature Reviews Cancer*, **2**, 19–27.

Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the

evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

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

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

A Implementations

Let us briefly outline the use of the SAS macro SURBINBIN that can be used to perform the analyses described in this paper. The macro is invoked in the following fashion:

```
%surbinbin(response=,endpoint=,trial=,subject=,data=,trt=,adj=,
initfix=,initd=,initcorr=,red=,type=,dmat=,outf=
,plot=,drive=,file=,solutionf=)
```

where

es response: Name of the response variable.

endpoint: Name of the endpoint indicator (-1 =surrogate endpoint, 1 =true endpoint).

trial: Name of the unit of analysis (center, trial,...).

subject: Name of the variable indicating the unique subject identification number.

data: Name of the input dataset. See the macro description on data formatting and layout for further detail.

trt: Name of the treatment indicator variable.

adj: A choice for using weighted (adj$= 1$)or unweighted(adj$= 0$) regression at the second stage, i.e., the choice made along the measurement error dimension.

red: A choice for using reduced (red$= 1$) or full (red$= 0$) model, i.e., whether or not the random intercept is taken into account when calculating R^2_{trial}.

initd: If type is set to 4, this string is used to enter initial values for the covariance matrix D of random effects. The elements of the lower-triangular part are to be listed by row.

initfix: If type is set to 4, this string is used to enter initial values for the fixed-effects parameters (mu_S, $beta_S$, mu_T, and $beta_T$).

crit: If type is set to 4, this string is used to enter the absolute parameter convergence criterion. The default value is 0.0001.

initcorr: If type is set to 4, this string is used to enter initial values for ρ, i.e., the residual correlation parameter.

type: A choice for using the different modeling approaches (1–4):

1. univariate fixed effects;
2. bivariate fixed effects;
3. univariate random effects;
4. bivariate random effects.

dmat: A choice for printing the matrix of the random terms to the output file (1 =yes, 0 =no).

solutionf: A choice for printing the solution for fixed effect to the output file (1 =yes, 0 =no).

outf: A choice for printing the trial specific random effects, for type$= 3$, or fixed effects, for type$= 1$ or 2, to the output file (1 =yes, 0 =no).

plot: A choice for printing the plot of the raw outcomes, residuals and treatment effects of the main endpoint against those of the surrogate endpoint (plot$= 1$).

drive: The drive on the computer where the user wants to save the output file (e.g., A, C, or D).

file: The name of the output file containing the macro results.

rescale: The option to control the size of the bubble plots when the treatment effects on the true endpoint are plotted against those of the surrogate endpoints. It can take integer values or fractions depending on the size of the plots. It is important that the endpoint indicator be coded as $-1/+1$, with 1 indicating the true endpoint.

Table 1: *Simulation study. Univariate models. Asymptotic confidence interval are calculated from the asymptotic sampling distribution of* R^2.

# trials	# subjects	R^2_{indiv}	bootstrap c.i.	asymptotic c.i.
Mixed-effects model: individual-level surrogacy				
5	10	0.2661	(0.0108;0.6816)	(0.0917;0.5070)
5	60	0.2365	(0.1292;0.3504)	(0.1516;0.3340)
30	10	0.3362	(0.1964;0.4583)	(0.2394;0.4406)
30	60	0.2388	(0.1823;0.2788)	(0.2022;0.2775)
5	100	0.2358	(0.1392;0.3489)	(0.1687;0.3104)
5	300	0.2211	(0.1354;0.3102)	(0.1822;0.2627)
30	100	0.2340	(0.1971;0.2816)	(0.2006;0.2502)
30	300	0.2220	(0.1839;0.2561)	(0.2054;0.2225)
Mixed-effects model: trial-level surrogacy				
5	10	0.7037	(0.0332;0.9949)	(0.3199;0.9293)
5	60	0.8880	(0.5288;0.9978)	(0.5531;0.9849)
30	10	0.7225	(0.4826;0.8965)	(0.5098;0.8674)
30	60	0.8414	(0.7109;0.9188)	(0.6761;0.9348)
5	100	0.9014	(0.5550;0.9959)	(0.5719;0.9901)
5	300	0.9092	(0.5418;0.9993)	(0.6014;0.9907)
30	100	0.8596	(0.7548;0.9397)	(0.7066;0.9436)
30	300	0.8767	(0.7977;0.9415)	(0.7344;0.9520)
Fixed-effects model: individual-level surrogacy				
5	10	0.2274	(0.0036;0.5809)	(0.0684;0.4663)
5	60	0.2203	(0.1263;0.3297)	(0.1379;0.3160)
30	10	0.2231	(0.0886;0.3534)	(0.1413;0.3177)
30	60	0.2149	(0.1662;0.2593)	(0.1797;0.2523)
5	100	0.2232	(0.1135;0.3360)	(0.1577;0.2966)
5	300	0.2183	(0.1185;0.3070)	(0.1796;0.2597)
30	100	0.2172	(0.1830;0.2635)	(0.1897;0.2461)
30	300	0.2148	(0.1768;0.2524)	(0.1962;0.2200)
Fixed-effects model: trial-level surrogacy				
5	10	0.74967	(0.2237;0.9994)	(0.3429;0.9572)
5	60	0.83546	(0.3234;0.9980)	(0.4714;0.9779)
30	10	0.63242	(0.3971;0.8232)	(0.3980;0.8109)
30	60	0.66659	(0.4111;0.8379)	(0.4408;0.8323)
5	100	0.8462	(0.3826;0.9966)	(0.4770;0.9799)
5	300	0.8864	(0.5697;0.9986)	(0.5570;0.9873)
30	100	0.6780	(0.4351;0.8713)	(0.4528;0.8403)
30	300	0.7639	(0.5283;0.9317)	(0.5673;0.8913)

Table 2: *Simulation study. Bivariate models. Asymptotic confidence interval are calculated from the asymptotic sampling distribution of* R^2.

# trials	# subjects	R^2_{trial}	bootstrap c.i.	asymptotic c.i.
Mixed-effects model: individual-level surrogacy				
5	10	0.8088		(0.1167;1.0000)
5	60	0.6993		(0.4511;0.9491)
30	10	0.6340		(0.3267;0.8577)
30	60	0.6349		(0.5445;0.7378)
5	100	0.6863		(0.4814;0.9044)
5	300	0.6650		(0.5204;0.7967)
30	100	0.6392		(0.5695;0.7095)
30	300	0.6339		(0.5833;0.6804)
Mixed-effects model: trial-level surrogacy				
5	10	0.9749		(0.6438;1.0000)
5	60	0.9349		(0.5049;1.0000)
30	10	0.9231		(0.6530;0.9999)
30	60	0.9142		(0.7970;0.9996)
5	100	0.9433		(0.5005;1.0000)
5	300	0.9325		(0.5021;1.0000)
30	100	0.9152		(0.7932;0.9947)
30	300	0.9082		(0.7729;0.9984)
Fixed-effects model: individual-level surrogacy				
5	10	0.8298		(0.0492;0.9999)
5	60	0.6976		(0.4737;0.9999)
30	10	0.8487		(0.5242;0.9999)
30	60	0.6755		(0.5984;0.7589)
5	100	0.6843		(0.5058;0.8273)
5	300	0.6976		(0.5570;0.7440)
30	100	0.6682		(0.6066;0.7222)
30	300	0.6481		(0.6163;0.6845)
Fixed-effects model: trial-level surrogacy				
5	10	0.7975	(0.2130;0.9995)	(0.3842;0.9699)
5	60	0.8427	(0.3645;0.9988)	(0.4844;0.9795)
30	10	0.5675	(0.3313;0.7811)	(0.3280;0.7654)
30	60	0.6834	(0.4700;0.8532)	(0.4602;0.8435)
5	100	0.8500	(0.3796;0.9965)	(0.4828;0.9802)
5	300	0.8893	(0.5515;0.9996)	(0.5633;0.9877)
30	100	0.7126	(0.4928;0.8711)	(0.4948;0.8624)
30	300	0.8059	(0.6373;0.9204)	(0.6244;0.9155)

Table 3: *Acute Migraine Study. Estimates (confidence intervals) for trial-level and individual-level surrogacy for the* **photophobia** *symptom.*

Trial-level surrogacy			
Fixed effects		Random effects	
Unweighted	Weighted	Unweighted	Weighted
Univariate approach			
0.7579	0.7579	0.8112	0.8886
(0.5712;0.8817)	(0.5712;0.8817)	(0.6367;0.9066)	(0.8134;0.9567)
Bivariate approach			
0.7336	0.7336	0.9587*	
(0.5426;0.8688)	(0.5426;0.8688)	(0.6966;1.000)	

Individual-level surrogacy	
Fixed effects	Random effects
Univariate approach (ITA based)	
0.5016	0.5885
(0.4354;0.5681)	(0.5221;0.6540)
Bivariate approach (probit, latent scale)	
0.8959	0.8664
(0.8822;0.9095)	(0.6042;1.000)

$*$: *This value is unreliable due to ill-conditioning of the variance-covariance matrix from which it was calculated.*

Table 4: *Acute Migraine Study. Estimates (confidence intervals) for trial-level and individual-level surrogacy for the* **phonophobia** *symptom.*

Trial-level surrogacy			
Fixed effects		Random effects	
Unweighted	Weighted	Unweighted	Weighted
Univariate approach			
0.5792	0.6311	0.6934	0.8119
(0.3638;0.7945)	(0.4340;0.8281)	(0.5218;0.8650)	(0.6979;0.9258)
Bivariate approach			
0.5591	0.6776	0.9379*	
(0.3375;0.7807)	(0.4992;0.8560)	(0.8092;1.000)	

Individual-level surrogacy	
Fixed effects	Random effects
Univariate approach (ITA based)	
0.3916	0.4853
(0.3281;0.4570)	(0.4176;0.5499)
Bivariate approach (probit, latent scale)	
0.8105	0.6828
(0.7868;0.8342)	(0.5543;1.000)

$*$: *This value is unreliable due to ill-conditioning of the variance-covariance matrix from which it was calculated.*

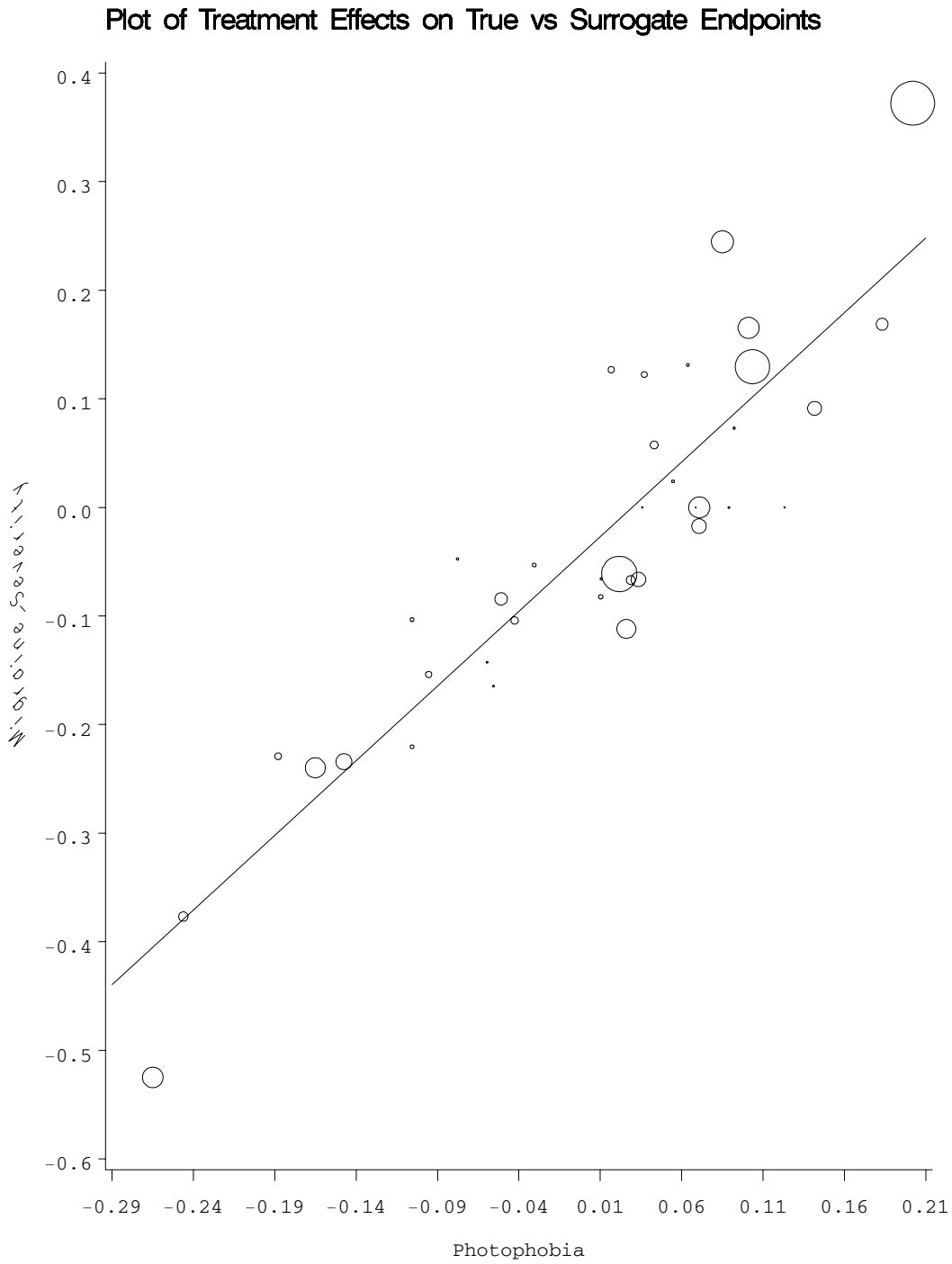

Figure 1: *Acute Migraine Study. Bubble plot of trial-specific treatment effect on the surrogate versus true endpoints. The horizontal axis represents treatment effect on the surrogate enpoint, whereas the vertical axis refers to treatment effect on the true endpoint.*