Challenges in the Methodology for the Validation of Surrogate Endpoints In Randomized Trials

Geert Molenberghs[1], Marc Buyse[2], Helena Geys[1], Didier Renard[1], Tomasz Burzykowski[1]

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium; E-mail: geert.molenberghs@luc.ac.be
[2] International Drug Development Institute, Brussels, Belgium

Abstract: The validation of surrogate endpoints has been studied by Prentice (1989). He presented a definition as well as a set of criteria that are equivalent if the surrogate and true endpoints are binary. Freedman (1992) supplemented these criteria with the so-called *proportion explained*. Buyse and Molenberghs (1998) proposed to replace the proportion explained by two quantities: (1) the *relative effect* linking the effect of treatment on both endpoints and (2) the *adjusted association*, an individual-level measure of agreement between both endpoints. In this paper, we argue that a meta-analytic approach should be adopted because it overcomes difficulties which necessarily surround validation efforts based on a single trial.

Keywords: Adjusted Assocation; Meta-analysis; Proportion Explained; Random-effects Model; Relative Effect; Surrogate Endpoint; Validation.

1 Introduction

Surrogate endpoints are referred to as endpoints that can replace or supplement other endpoints in the evaluation of experimental treatments or other interventions. For example, surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the endpoints of interest, which are referred to as the "true" endpoints (Ellenberg and Hamilton 1989). Prentice (1989) proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated. This framework was extended by Freedman's *proportion explained* Freedman *et al* 1992). Buyse and

Molenberghs (1998) proposed the *relative effect* and *adjusted association*. All of these concepts are developed within the context of a single trial. In this paper, we will show how they exhibit some fundamental problems that can be overcome when shifting to a multiple-trial framework.

The paper starts with the case of a single trial (Section 2). In Section 2 we briefly review Prentice's definition and criteria. The proportion explained is introduced in Section 2.1, and the relative effect and adjusted association in Section 2.2. Problems with these concepts are discussed. The second part of the paper is devoted to the case of multiple trials (Section 3), with further illustration of the problems surrounding the proportion explained.

2 Data from a Single Trial

Throughout the paper, we will adopt the following notation: T and S are random variables that denote the true and surrogate endpoints, respectively, and Z is an indicator variable for treatment. For ease of exposition, we will assume that S and T are normally distributed. The effect of treatment on S and T can be modelled as follows:

$$S_i|Z_i = \mu_S + \alpha Z_i + \varepsilon_{Si}, \tag{1}$$

$$T_i|Z_i = \mu_T + \beta Z_i + \varepsilon_{Ti}, \tag{2}$$

where $i = 1, \ldots, n$ indicates patients, and the error terms have a joint zero-mean normal distribution with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{3}$$

In addition, the relationship between S and T can be described by a regression of the form

$$T_i|S_i = \mu + \gamma S_i + \varepsilon_i. \tag{4}$$

Prentice (1989, p. 432) proposed to define a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint".

Prentice derived operational criteria that are equivalent to his definition. These criteria require that: (1) treatment has a significant impact on the surrogate endpoint; (2) treatment has a significant impact on the true endpoint; (3) the surrogate endpoint has a significant impact on the true endpoint; and (4) the full effect of treatment upon the true endpoint is captured by the surrogate. The last criterion is verified through the conditional distribution of the true endpoint, given treatment *and* surrogate endpoint, derived from (1)–(2):

$$T_i|Z_i, S_i = \tilde{\mu}_T + \beta_S Z_i + \gamma_Z S_i + \tilde{\varepsilon}_{Ti}, \tag{5}$$

where

$$\beta_S = \beta - \sigma_{TS}\sigma_{SS}^{-1}\alpha, \quad (6)$$

$$\gamma_Z = \sigma_{TS}\sigma_{SS}^{-1}, \quad (7)$$

and the variance of $\tilde{\varepsilon}_{Ti}$ is given by

$$\sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}. \quad (8)$$

It is usually stated that the fourth criterion requires that the parameter β_S be equal to zero (see also Section 2.2). Buyse and Molenberghs (1998) showed that the last two criteria are necessary and sufficient for binary responses, but not in general.

2.1 The Proportion Explained

Freedman *et al* (1992) argued that the last Prentice criterion raises a conceptual difficulty since it requires the statistical test for treatment effect on the true endpoint to be *non*-significant after adjustment for the surrogate and proposed to calculate the proportion of the treatment effect mediated by the surrogate:

$$PE = \frac{\beta - \beta_S}{\beta},$$

with β_S and β obtained respectively from (5) and (2). In this paradigm, a valid surrogate would be one for which the proportion explained (PE) is equal to one. In practice, a surrogate would be deemed acceptable if the lower limit of its confidence interval of PE was "sufficiently" large.

Some difficulties surrounding the PE have been described in the literature (Buyse and Molenberghs 1998, Daniels and Hughes 1997, Volberding *et al* 1990, Choi *et al* 1993, Lin *et al* 1997, Flandre and Saidi 1999). PE will tend to be unstable when β is close to zero, a situation that is likely to occur in practice. The confidence limits of PE will tend to be rather wide (and sometimes even unbounded if Fieller confidence intervals are used), unless large sample sizes are available or a very strong effect of treatment on the true endpoint is observed. Another complication arises when (5) is not the correct conditional model, and an interaction term between Z_i and S_i needs to be included. In that case, defining the PE becomes problematic.

2.2 The Relative Effect and Adjusted Association

Buyse and Molenberghs (1998) suggested to calculate another quantity for the validation of a surrogate endpoint: the relative effect (RE), which is the ratio of the effects of treatment upon the final and the surrogate endpoint. Formally:

$$RE = \beta/\alpha. \quad (9)$$

They also considered the treatment-adjusted association between the surrogate and the true endpoint: $\rho_Z = \sigma_{ST}/\sqrt{\sigma_{SS}\sigma_{TT}}$. Now, a simple relationship can be derived between PE, RE, and ρ_Z. Let us define $\lambda^2 = \sigma_{TT}\sigma_{SS}^{-1}$. It follows that $\lambda\rho_Z = \sigma_{ST}\sigma_{SS}^{-1}$ and, from (6), $\beta_S = \beta - \rho_Z\lambda\alpha$. As a result, we obtain

$$PE = \lambda\rho_Z\frac{\alpha}{\beta} = \lambda\rho_Z\frac{1}{RE}. \tag{10}$$

2.3 Problems With Single-Trial Measures

Expression (10) allows us to make several useful observations. It is clear from (10) that the PE is *not* a proportion. Indeed, each of λ and RE can take values over the entire real line. Further, it is hard to interpret PE because it amalgamates three sources of information: (1) the adjusted assocation ρ_Z, which is a measure of association between the surrogate and the true endpoints *at the individual level*; (2) the RE, which expresses the relationship between the treatment effects on the surrogate and the true endpoint *at the trial level*; (3) the variance ratio λ^2, which is a nuisance parameter, not to be viewed as a useful validation measure.

The fact that the PE is ill defined, except in trivial cases, and the relationship between the three measures introduced above, will be studied by means of three thought experiments. The first two experiments concentrate on "perfect" conditions, while the last one focuses on general conditions.

Thought Experiment 1. The PE is obviously equal to one in simple situations of perfect surrogacy, for instance if T is linearly related to S ($T = aS + b$), for then (1) and (2) can be rewritten as

$$S_i|Z_i = \mu_S + \alpha Z_i + \varepsilon_{Si}, \tag{11}$$
$$T_i|Z_i = b + a\mu_S + a\alpha Z_i + a\varepsilon_{Si}, \tag{12}$$

and obviously $\rho_Z = 1$, $\lambda = a$ and $RE = a$.

However, it is possible to construct examples where PE can be chosen to take any arbitrary (positive) value, depending on the values of ρ_Z, λ and RE. To this end we conduct two further thought experiments.

Thought Experiment 2. Assume $\rho_Z = 1$ and $RE = 1$, and suppose further that we could reduce (increase) the variance of the surrogate endpoint while keeping all other quantities unaffected, say by improving (deteriorating) the precision of its measurement. Then, (1)–(2) would become

$$S_i|Z_i = \mu_S + \alpha Z_i + \varepsilon_{Si}, \tag{13}$$
$$T_i|Z_i = \mu_S + \alpha Z_i + \lambda\varepsilon_{Si}. \tag{14}$$

λ is arbitrary and hence so is PE, despite the fact that (13)–(14) describe a very desirable situation. The key behind this somewhat artificial and counterintuitive

thought experiment is that the systematic components are kept constant, the random error terms are in *perfect* correlation. Then, knowledge about the surrogate endpoint enables exact prediction of the true endpoint: $E[T_i|Z_i, S_i] = T_i$. Now, we would like to call the situation described by (13)–(14) "perfect", even though PE may not be equal to one, nor β_S equal to zero. This casts doubts on the fourth Prentice criterion, which states that the full effect of treatment should be captured by the surrogate, even though this criterion has much intuitive appeal. In the above example, the conditional distribution of the true endpoint, given treatment and surrogate endpoint, is

$$T_i|Z_i, S_i = \tilde{\mu}_T + \alpha(1-\lambda)Z_i + \lambda S_i, \tag{15}$$

which shows that the true endpoint does depend on treatment, although the residual, unexplained, variability in the true endpoint has been eliminated. Then, (8) vanishes, which is equivalent to stating that $\rho_Z = 1$. This suggests to focus on the adjusted association, rather than on the adjusted treatment effect upon the true endpoint. Note that perfection in this context has no implication for the surrogate *across trials*. To study the latter very important quality it is necessary to turn to RE or even to a multi-trial setting (Section 3).

Thought Experiment 3. We will now switch to general conditions and consider two transformations of the surrogate endpoint:

$$S_i^{(1)} = \phi S_i + \psi = (\phi\mu_S + \psi) + \phi\alpha Z_i + \phi\varepsilon_{Si}, \tag{16}$$

$$S_i^{(2)} = \mu_S + \alpha Z_i + \phi\varepsilon_{Si}. \tag{17}$$

Note that the second transformation cannot be conducted through a simple transformation of a dataset variable. It might refer, for example, to a situation in a sequence of trials where at some point the precision changed due to a change in instrument or method used to measure the surrogate.

Transformation (16) operates on the fixed and random parts of the surrogate endpoint alike whereas transformation (17) operates on the random part only. The second transformation is similar to one in the second thought experiment, except that we now consider the general rather than the perfect situation. It is easy to show that the following relationships hold between the validation measures:

$$RE^{(1)} = RE/\phi, \qquad \rho_Z^{(1)} = \rho_Z, \qquad \lambda^{(1)} = \lambda/\phi, \qquad PE^{(1)} = PE,$$

$$RE^{(2)} = RE, \qquad \rho_Z^{(2)} = \rho_Z, \qquad \lambda^{(2)} = \lambda/\phi, \qquad PE^{(2)} = PE/\phi,$$

with obvious notation. Thus, for transformation (16) there is no impact on the PE, but under (17), PE is rescaled with an arbitrary amount.

There are also problems with the RE. Indeed, while the adjusted association expresses agreement between both endpoints at the individual level, the trialist will want to know how the *trial-specific* treatment effect on T can be predicted from the treatment effect on S. RE serves this purpose, but it is typically based on information from only one trial. It might not be constant for all trials testing the therapeutic question under consideration. The constancy of RE implies that

the relation between α and β is linear through the origin. This assumption may be untenable in practice, and it cannot be verified from a single trial. Therefore, it will prove useful to adopt an alternative definition of surrogacy based on a meta-analysis of several trials.

3 Data from Several Trials

Using ideas from Buyse *et al* (2000), we now extend the setting and notation by supposing we have data from $i = 1, \ldots, N$ trials, in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. We denote the true and surrogate endpoints and the treatment indicator by T_{ij}, S_{ij}, and Z_{ij}, respectively.

Linear models (1) and (2) can be rewritten as:

$$S_{ij}|Z_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{18}$$

$$T_{ij}|Z_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{19}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z on the endpoints in trial i, and ε_{Si} and ε_{Ti} are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix (3), as before. Due to the replication at the trial level, we can impose a distribution on the trial-specific parameters with mean $(\mu_S, \mu_T, \alpha, \beta)^T$ and covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{20}$$

These authors introduced trial-level and individual-level measures of surrogacy:

$$R^2_{\text{trial}} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{21}$$

and

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

We have argued at the end of Section 2 that, while the concept behind the fourth Prentice criterion has intuitive appeal, it is not captured by the PE. We also argued that RE is based on too strong assumptions to be useful. Having introduced measures of surrogacy at the trial-level and at the individual-level, it is now possible to explore these issues further.

The proportion explained (10), derived in Section 2.1 for the single-trial case, can be calculated for each trial within the meta-analysis:

$$PE_i = \lambda \rho_Z \frac{1}{RE_i}, \tag{22}$$

where $RE_i = \beta_i/\alpha_i$.

Let us now examine how the PE_i behaves relative to the R^2 measures. To make the point clearly, it is useful to concentrate on a "perfect" surrogate, i.e., one for which $R^2_{\text{trial}} = 1$ and $R^2_{\text{indiv}} = \rho_Z^2 = 1$.

Perfect Surrogate at the Trial Level. Let us first assume that the surrogate is perfect at the trial level, i.e., $R^2_{\text{trial}} = 1$. Then the relationship between α_i and β_i is deterministic, and (22) becomes

$$PE_i = \rho_Z \lambda \frac{\alpha_i}{\theta_0 + \theta_a \alpha_i + \theta_m \mu_{Si}}, \tag{23}$$

with obvious notation. Thus, even if the important condition $R^2_{\text{trial}} = 1$ is satisfied, and one can predict the treatment effect on the true endpoint without error from the treatment effect on the surrogate endpoint, PE_i cannot be constant across trials, and consequently would not be equal to unity in all of them. Note that also RE_i is not constant across trials. The reason is that for RE_i to be constant the relationship between α_i and β_i must be multiplicative.

Perfect Surrogate at the Individual Level. Let us now make the additional assumption that the surrogate is also perfect at the individual level, i.e., $\rho_Z = 1$.

In this case, (23) becomes

$$PE_i = \lambda \frac{\alpha_i}{\theta_0 + \theta_a \alpha_i + \theta_m \mu_{Si}}. \tag{24}$$

and the property of non-constant PE_i and RE_i persists, again due to the linear but non-multiplicative relationship between α_i and β_i.

Constant Relative Effect. Let us make the final assumption that a simple multiplicative relationship holds between α_i and β_i, i.e., $\theta_0 = \theta_m = 0$ and hence $RE_i = \theta_a$. Thus,

$$PE = PE_i = \frac{\lambda}{\theta_a}. \tag{25}$$

Now, RE_i is constant and so is PE_i, but the latter is still a function of two quantities: (1) the multiplicative factor θ_a linking the treatment effects in each trial and (2) the multiplicative factor λ linking the two error terms in each patient.

Clearly, under the three asumptions made above, the surrogate and true endpoints are identical, up to scaling factors that translate the treatment effects within a trial and the subject-specific deviations within each patient. Yet, depending on the values of θ_a and λ, the PE can assume any positive real value.

4 Discussion

In this note, we have argued that a classical approach to surrogate marker validation, based on the Prentice criteria and measures derived therefrom, such as the

proportion explained and the relative effect, is surrounded with difficulties. We have argued a meta-analytic framework is both more elegant and more principled. Meta-analytic developments, similar to the ones done here for normal outcomes, have been done for binary, survival, and longitudinal outcomes, and for situations where the true and surrogate outcomes are of a different type.

References

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.

Choi, S., Lagakos, S., Schooley, R.T., and Volberding, P.A. (1993) CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Annals of Internal Medicine*, **118**, 674–680.

Daniels, M.J. and Hughes, M.D. (1997) Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1515–1527.

Ellenberg, S.S. and Hamilton, J.M. (1989) Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine*, **8**, 405–413.

Flandre, P. and Saidi, Y. (1999) Letters to the editor: estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, **18**, 107–115.

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

Lin, D.Y., Fleming, T.R., and DeGruttola, V. (1997) Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, **16**, 1515–1527.

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

Volberding, P.A., Lagakos, S.W., Koch, M.A., *et al* (1990) Zidovudine in asymptomtic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *New England Journal of Medicine*, **322**, 941–949.