J. R. Statist. Soc. A (2005)
168, *Part* 4, *pp.* 753–762

Resolving paradoxes involving surrogate end points

Stuart G. Baker, Grant Izmirlian and Victor Kipnis

National Cancer Institute, Bethesda, USA

[Received February 2004. Revised December 2004]

Summary. We define a surrogate end point as a measure or indicator of a biological process that is obtained sooner, at less cost or less invasively than a true end point of health outcome and is used to make conclusions about the effect of an intervention on the true end point. Prentice presented criteria for valid hypothesis testing of a surrogate end point that replaces a true end point. For using the surrogate end point to estimate the predicted effect of intervention on the true end point, Day and Duffy assumed the Prentice criterion and arrived at two paradoxical results: the estimated predicted intervention effect by using a surrogate can give more precise estimates than the usual estimate of the intervention effect by using the true end point and the variance is greatest when the surrogate end point perfectly predicts the true end point. Begg and Leung formulated similar paradoxes and concluded that they indicate a flawed conceptual strategy arising from the Prentice criterion. We resolve the paradoxes as follows. Day and Duffy compared a surrogate-based estimate of the effect of intervention on the true end point with an estimate of the effect of intervention on the true end point that uses the true end point. Their paradox arose because the former estimate assumes the Prentice criterion whereas the latter does not. If both or neither of these estimates assume the Prentice criterion, there is no paradox. The paradoxes of Begg and Leung, although similar to those of Day and Duffy, arise from ignoring the variability of the parameter estimates irrespective of the Prentice criterion and disappear when the variability is included. Our resolution of the paradoxes provides a firm foundation for future meta-analytic extensions of the approach of Day and Duffy.

Keyword: Prentice criterion

1. Introduction

We define a surrogate end point as an end point that is obtained sooner, at less cost or less invasively than a true end point and is used to make conclusions about the effect of intervention on the true end point. Examples include a stage of cancer as a surrogate end point for death from cancer or diastolic blood pressure as a surrogate end point for strokes. In the context of randomized trials (which is the focus here), the objective is to use the surrogate end point to make inference about the effect of an intervention on the true end point in an application trial in which only the surrogate end point is observed. Before the use of the surrogate end point in an application trial, it must be validated by using data from a trial with both surrogate and true end points, which we call a validation trial.

Much early statistical work on surrogate end points focused on using the surrogate end point to *replace* the true end point. Because the surrogate and true end points are on different scales a direct comparison is meaningless. Therefore inference was confined to hypothesis testing. In the situation of hypothesis testing, validation consists of showing that rejection of the null hypothesis under the surrogate end point implies rejection of the null hypothesis under the true end point in a validation trial. In a landmark paper, Prentice (1989) gave criteria when the null

Address for correspondence: Stuart G. Baker, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, EPN 3131, 6130 Executive Boulevard MSC 7354, Bethesda, MD 20892-7354, USA.
E-mail: sb16i@nih.gov

 0964–1998/05/168753

hypothesis under the true end point implies the null hypothesis under the surrogate end point (so that rejecting the null hypothesis under the surrogate end point implies rejecting the null hypothesis under the true end point). The key criterion, called the Prentice criterion by Begg and Leung (2000), is that the distribution of the true end point conditional on the surrogate end point does not also depend on the randomization group. Rejecting the Prentice criterion indicates that the surrogate end point cannot validly replace the true end point. However, not rejecting the Prentice criterion does not mean that the surrogate end point is validated. Consequently, various summary measures (such as the proportion of treatment effect explained) have been proposed when the Prentice criterion cannot be rejected (e.g. Freedman *et al.* (1992)). A general difficulty with interpreting many of these summary measures is that they are not directly linked to the effect of intervention on the true end point and, as a consequence, it is difficult to specify a target value that indicates a validated surrogate end point.

Other statistical work on surrogate end points has focused on using a surrogate end point to *predict* the effect of intervention on the true end point. The basic idea is to use data from one or more previous trials to construct a model relating surrogate to true end points. The model is then applied to data on the surrogate end point in the validation trial to predict the effect of intervention on the true end point in the validation trial. (In a more complicated and more informative variation of this approach, each trial from a set of previous trials is successively selected as the validation trial and the remaining trials are used to fit the model. For simplicity here, we consider only a single validation trial.) In this context, one can validate the surrogate end point by comparing estimates and confidence intervals for

(a) the predicted effect of intervention on the true end point in the validation trial based on the model and the surrogate end point in the validation trial and
(b) the direct estimate of the effect of intervention on the true end point.

If the confidence intervals are similar (where the degree of similarity depends on the application), we say that the surrogate end point is validated and one can have more confidence in using the surrogate end point to predict the effect of intervention on the true end point in the application trial in which only the surrogate end point is observed.

Various types of model have been proposed for relating surrogate and true end points in one or more previous trials. Some models have involved trial level statistics for surrogate and true end points in multiple previous trials (e.g. Gail *et al.* (2000) and Buyse *et al.* (2000)). Other models have involved individual level associations between surrogate and true end points, as proposed by Morrison (1991) and Day and Duffy (1996) with a single previous trial. We find this latter approach appealing for binary surrogate and true end points because no additional assumptions are needed for modelling the association between surrogate and true end points, unlike the situation with continuous surrogate and true end points. Current research involves meta-analytic extensions of this approach with a random-effects component to capture variability between trials. However, it is necessary to resolve clearly the paradoxical results of Day and Duffy (1996) that, if not addressed, could undermine the foundations for this proposed meta-analytic approach.

Day and Duffy (1996) explicitly incorporated the Prentice criterion in their model and obtained the paradoxical results that Begg and Leung (2000) attributed to a conceptual flaw related to the Prentice criterion. We think that part of the confusion over these paradoxes is the common mistaken belief that the Prentice criterion, which was proposed for valid hypothesis testing, is necessary for valid estimation of the effect of intervention on the true end point based on data from the surrogate end point. Baker and Kramer (2003) showed graphically that this was not so. In particular, they showed that

(a) the Prentice criterion corresponds to identical lines depicting the relationship of surrogate to true end point in each randomization groups and
(b) predicting the effect of intervention on the true end point requires only correct prediction of the aforementioned lines regardless of whether or not they are identical.

In Sections 2.3 and 3.2, we present formulae for predicting the effect of intervention on the true end point that do not require the Prentice criterion.

The paper is organized as follows. In Section 2 we describe how the paradox in Day and Duffy (1996) arises because the Prentice criterion is assumed in only one of two estimates being compared. In Section 3, we explain how the paradoxes in Begg and Leung (2000), which are similar to those in Day and Duffy (1996), arise for a different reason. We conclude that the method of using the surrogate end point to predict the effect of intervention on the true end point is *not* conceptually flawed.

2. Explaining the paradoxes in Day and Duffy (1996)

Day and Duffy (1996) compared the following two variances for the estimated effect of intervention on the true end point in a validation trial with both surrogate and true end points:

(a) the variance of the estimated predicted effect of intervention on the true end point based on the surrogate end point and
(b) the variance of the estimated effect of intervention on the true end point based on the true end point.

In Section 2.1, we show that the paradoxes of Day and Duffy (1996) arise when the Prentice criterion is assumed for (a) but not (b). In Section 2.2, we show that there is no paradox when both (a) and (b) assume the Prentice criterion. In Section 2.3, we show that there is no paradox when neither (a) nor (b) assumes the Prentice criterion.

2.1. Prentice criterion for only surrogate-based estimate

We present the approach of Day and Duffy (1996) with a different notation to improve the clarity. Let T denote the true end point of cancer death ($T=1$) or not ($T=0$), S denote the surrogate end point which is the category of tumour size at cancer detection, Z denote the intervention group ($Z=1$) or control group ($Z=0$) and J denote the validation trial ($J \equiv$ validation) or the previous trial ($J \equiv$ previous).

Day and Duffy (1996) assumed that the Prentice criterion is satisfied, namely $\text{pr}(T=t|S=s, Z=z, J=j)=\text{pr}(T=t|S=s, J=j)$. Let

$$\theta_s = \text{pr}(T=1|S=s, J \equiv \text{validation}),$$

$$\pi_{zs} = \text{pr}(S=s|Z=z, J \equiv \text{validation}),$$

$$\theta_s^* = \text{pr}(T=1|S=s, J \equiv \text{previous}),$$

$$\pi_{zs}^* = \text{pr}(S=s|Z=z, J \equiv \text{previous}).$$

Also let n_{zst} denote the number of subjects in the validation trial who are in group z with surrogate s (even if not observed) and true end point t. Day and Duffy (1996) excluded from the analysis subjects with no tumour detected and assumed that the same numbers of tumours are detected in both randomization groups, namely $N=n_{0++}=n_{1++}$, where '+' in a subscript indicates summation over the corresponding index. Because the number of subjects with a tumour

is small, we assume that n_{zst} follows a Poisson distribution. Therefore $n_{zs+} \sim \text{Poisson}(N\pi_{zs})$, $n_{z+1} \sim \text{Poisson}(N\Sigma_s \pi_{zs}\theta_s)$ and $n_{zs1}|n_{zs+} \sim \text{binomial}(\theta_s, n_{zs+})$. The end point that is of interest is the difference in the logarithm of expected numbers of cancer deaths among subjects with a tumour in each group: $\Delta = \log\{E(n_{1+1})\} - \log\{E(n_{0+1})\}$. For mathematical convenience we write $\Delta = \log\{E(n_{1+1})/N\} - \log\{E(n_{0+1})/N\}$.

Under the model with the Prentice criterion we can write the effect intervention on the observed true end point as

$$\Delta^{\text{PC}}_{\text{obs}} = \log\Big(\sum_s \theta_s \pi_{1s}\Big) - \log\Big(\sum_s \theta_s \pi_{0s}\Big), \tag{1}$$

where the subscript 'obs' refers to the fact that the true end point is observed and the superscript 'PC' indicates Prentice criterion. Now consider the situation that is of interest which uses the surrogate end point in the *validation trial* and both the surrogate and the true end points in a *previous* trial. We define the predicted intervention effect as

$$\Delta^{\text{PC}}_{\text{pred}} = \log\Big(\sum_s \theta^*_s \pi_{1s}\Big) - \log\Big(\sum_s \theta^*_s \pi_{0s}\Big), \tag{2}$$

where θ^*_s, the parameter from the previous trial, substitutes in equation (1) for θ_s, the parameter from the validation trial. Let n^*_{zst} denote the counts in the previous trial and let $N^* = n^*_{0++} = n^*_{1++}$. We assume that n^*_{zst} follows a Poisson distribution, so $(n^*_{zs1}, n^*_{zs0}|n^*_{+s+}) \sim \text{binomial}(\theta^*_s, n^*_{+s+})$ and $n^*_{+s+} \sim \text{Poisson}(N^*\pi^*_{+s})$. The estimated predicted intervention effect is

$$\hat{\Delta}^{\text{PC}}_{\text{pred}} = \log\Big(\sum_s \hat{\theta}^*_s \hat{\pi}_{1s}\Big) - \log\Big(\sum_s \hat{\theta}^*_s \hat{\pi}_{0s}\Big), \tag{3}$$

where $\hat{\theta}^*_s = n^*_{+s1}/n^*_{+s+}$ and $\hat{\pi}_{zs} = n_{zs+}/N$. Because $\text{var}(\hat{\pi}_{zs}) = \pi_{zs}/N$ and

$$\begin{aligned}
\text{var}(\hat{\theta}^*_s) &= E\{\text{var}(\hat{\theta}^*_s|n^*_{+s+})\} + \text{var}\{E(\hat{\theta}^*_s|n^*_{+s+})\} \\
&= E\left\{\frac{\theta^*_{zs}(1-\theta^*_{zs})}{n^*_{+s+}}\right\} \\
&\approx \frac{\theta^*_{zs}(1-\theta^*_{zs})}{N^*\pi^*_{+s}},
\end{aligned} \tag{4}$$

the asymptotic variance of $\hat{\Delta}^{\text{PC}}_{\text{pred}}$ under the delta method is

$$\text{var}(\hat{\Delta}^{\text{PC}}_{\text{pred}}) = \frac{1}{N}\sum_z \frac{\sum_s \theta^{*2}_s \pi_{zs}}{\big(\sum_s \theta^*_s \pi_{zs}\big)^2} + \frac{1}{N^*}\sum_s \frac{\theta^*_s(1-\theta^*_s)}{\pi^*_{+s}}\left(\frac{\pi_{1s}}{\sum_s \theta^*_s \pi_{1s}} - \frac{\pi_{0s}}{\sum_s \theta^*_s \pi_{0s}}\right)^2. \tag{5}$$

Implicitly assuming that $\theta^*_s = \theta_s$ in a numerical example, Day and Duffy (1996) found that the second term in equation (5) was negligible. In their calculations, Day and Duffy (1996) specified a previous trial with $N^* > N$ and $\text{var}(\hat{\theta}^*_s) < \text{var}(\hat{\theta}_s)$. To isolate the effect of the Prentice criterion, we recomputed the second term in equation (5) assuming that $N^* = N$ and $\text{var}(\hat{\theta}^*_s) = \text{var}(\hat{\theta}_s)$ and found that the second term was still negligible. Dropping the second term in equation (5), Day and Duffy (1996) obtained the following approximate asymptotic variance:

$$\text{var}(\hat{\Delta}^{\text{PC}}_{\text{pred}}) \approx \frac{1}{N}\sum_z \frac{\sum_s \theta^{*2}_s \pi_{zs}}{\big(\sum_s \theta^*_s \pi_{zs}\big)^2}. \tag{6}$$

For comparison, Day and Duffy (1996) estimated the observed intervention effect by

$$\hat{\Delta}_{\text{obs}}^{\text{DD}} = \log(n_{1+1}/N) - \log(n_{0+1}/N) \tag{7}$$

with asymptotic variance

$$\text{var}(\hat{\Delta}_{\text{obs}}^{\text{DD}}) = \sum_z \frac{\text{var}(n_{z+1})}{E(n_{z+1})^2} = \frac{1}{N} \sum_z \frac{1}{\sum_s \theta_s \pi_{zs}}, \tag{8}$$

where the superscript 'DD' indicates the Day and Duffy model. Importantly the Prentice criterion is not invoked in equation (7) and only appears peripherally in equation (8) in the formula for expected counts. Comparing equation (6) with equation (8), Day and Duffy (1996) obtained the following paradoxes.

(a) *Paradox 1*: the variance of the estimated predicted intervention effect in equation (8) is smaller than the variance of the estimated observed intervention effect in equation (6).
(b) *Paradox 2*: if the surrogate end point perfectly predicts the true end point, i.e. $\theta_s = 1$ or $\theta_s = 0$, the variance of the estimated predicted intervention effect is the largest.

Day and Duffy explained paradox 1 as follows:

> 'the surrogates usually provide us with more information per subject than the true end point, in this case a probability of death rather than the binary observation of death or no death'.

We think that a better explanation is that

(a) the estimate when using the surrogate end point to predict the true end point in equation (3) assumes the Prentice criterion whereas
(b) the estimate with the observed true end point in equation (7) does not.

Because (a) postulates a more parsimonious model, the estimate has a smaller variance than in (b).

We agree with the explanation of paradox 2 in Day and Duffy (1996):

> 'The true end point is usually also influenced by numerous other factors which, given the surrogate, are not further affected by treatment. Bias with respect to these is controlled for by randomization, but they add random error to the true end point.'

We would even state that under the Prentice criterion the surrogate end point is the *de facto* end point of interest and the true end point only adds noise.

2.2. Prentice criterion for both estimates

To check our conclusions about paradox 1, suppose that we invoke the Prentice criterion for estimating the effect of intervention on the true end point. Instead of equation (7) the estimated observed effect of intervention on the true end point is $\hat{\Delta}_{\text{obs}}^{\text{PC}} = \log(\Sigma_s \hat{\theta}_s \hat{\pi}_{1s}) - \log(\Sigma_s \hat{\theta}_s \hat{\pi}_{0s})$ with asymptotic variance given by equation (5) with $\theta_s^* = \theta_s$, $\pi_{+s}^* = \pi_{+s}$ and $N^* = N$. Suppose that the previous trial has the same size and the same distribution of surrogate and true end points as the validation trial so again $\theta_s^* = \theta_s$, $\pi_{+s}^* = \pi_{+s}$ and $N^* = N$. Then $E(\hat{\Delta}_{\text{pred}}^{\text{PC}}) = E(\hat{\Delta}_{\text{obs}}^{\text{PC}})$ and $\text{var}(\hat{\Delta}_{\text{pred}}^{\text{PC}}) = \text{var}(\hat{\Delta}_{\text{obs}}^{\text{PC}})$. This result is sensible because the same information is used to estimate $\Delta_{\text{pred}}^{\text{PC}}$ and $\Delta_{\text{obs}}^{\text{PC}}$. Thus there is no paradox under this scenario.

2.3. Prentice criterion for neither estimate

As another check of our assertion that assuming the Prentice criterion for equation (3) but not for equation (7) leads to paradox 1, suppose that the Prentice criterion does not hold for either

the observed or the predicted effect of intervention on the true end point. Let

$$\theta_{zs} = \mathrm{pr}(T=1|S=s, Z=z, J \equiv \text{validation}),$$

$$\theta^*_{zs} = \mathrm{pr}(T=1|S=s, Z=z, J \equiv \text{previous}).$$

If the Prentice criterion is not assumed, the observed and predicted intervention effects are

$$\Delta_{\mathrm{obs}} = \log\Big(\sum_s \theta_{1s}\pi_{1s}\Big) - \log\Big(\sum_s \theta_{0s}\pi_{0s}\Big)$$

and

$$\Delta_{\mathrm{pred}} = \log\Big(\sum_s \theta^*_{1s}\pi_{1s}\Big) - \log\Big(\sum_s \theta^*_{0s}\pi_{0s}\Big) \tag{9}$$

respectively. From equation (9), the estimated predicted intervention effect without the Prentice criterion is

$$\hat{\Delta}_{\mathrm{pred}} = \log\Big(\sum_s \hat{\theta}^*_{1s}\hat{\pi}_{1s}\Big) - \log\Big(\sum_s \hat{\theta}^*_{0s}\hat{\pi}_{0s}\Big), \tag{10}$$

where $\hat{\theta}^*_{zs} = n^*_{zs1}/n^*_{zs+}$. The approximate asymptotic variance of the estimated predicted intervention effect is

$$\mathrm{var}(\hat{\Delta}_{\mathrm{pred}}) \approx \frac{1}{N}\sum_z \frac{\sum_s \theta^{*2}_{zs}\pi_{zs}}{\big(\sum_s \theta^*_{zs}\pi_{zs}\big)^2} + \frac{1}{N^*}\sum_z \frac{\sum_s \{\theta^*_{zs}(1-\theta^*_{zs})/\pi^*_{zs}\}\pi^2_{zs}}{\big(\sum_s \theta^*_{zs}\pi_{zs}\big)^2}. \tag{11}$$

Because equation (9) corresponds to a saturated model, the estimated observed intervention effect is

$$\hat{\Delta}_{\mathrm{obs}} = \log(n_{1+1}/N) - \log(n_{0+1}/N), \tag{12}$$

with asymptotic variance

$$\mathrm{var}(\hat{\Delta}_{\mathrm{obs}}) = \sum_z \frac{\mathrm{var}(n_{z+1})}{E(n_{z+1})^2} = \frac{1}{N}\sum_z \frac{1}{\sum_s \theta_{zs}\pi_{zs}}. \tag{13}$$

Suppose that the previous trial has the same size and the same distribution of surrogate and true end points as in the validation trial. In this case $N^* = N$, $\theta^*_{zs} = \theta_{zs}$ and $\pi^*_{zs} = \pi_{zs}$, so $E(\hat{\Delta}_{\mathrm{pred}}) = E(\hat{\Delta}_{\mathrm{obs}})$ and $\mathrm{var}(\hat{\Delta}_{\mathrm{pred}}) = \mathrm{var}(\hat{\Delta}_{\mathrm{obs}})$, which is what we would expect as the information content is the same for estimating the predicted and observed intervention effects. Again there is no paradox.

Interestingly, if the previous trial has a larger size but the same distribution of surrogate and true end points as in the validation trial $E(\hat{\Delta}_{\mathrm{pred}}) = E(\hat{\Delta}_{\mathrm{obs}})$ but $\mathrm{var}(\hat{\Delta}_{\mathrm{pred}}) \leqslant \mathrm{var}(\hat{\Delta}_{\mathrm{obs}})$. Thus, without assuming the Prentice criterion, it is possible for the estimated predicted intervention effect based on the surrogate to have a smaller variance than the observed intervention effect. However, this result differs from the paradox of Day and Duffy (1996).

3. Why the paradoxes in Begg and Leung (2000) differ

In the course of investigating the variance paradoxes in Day and Duffy (1996), Begg and Leung (2000) derived similar paradoxes and concluded that 'the conceptual strategy is flawed, and that the fundamental problem is the Prentice criterion'. As we shall show, we believe that the

paradoxes in Begg and Leung (2000) arise when one does not account for the variability of the parameter estimates and hence have a different explanation from those of the paradoxes in Day and Duffy (1996), even though the two sets of paradoxes appear similar.

3.1. Derivation of Begg and Leung (2000)

Begg and Leung (2000) presented the linear model $T_{zi} = \alpha_z + \beta_z S_{zi} + \varepsilon_{zi}$, where S_{zi} and T_{zi} (with realization s_{zi} and t_{zi}) are continuous surrogate and true end points for individual i in intervention group z in the validation trial. Under the model $E(\varepsilon_{zi}) = 0$, $\mathrm{var}(\varepsilon_{zi}) = \sigma_z^2$ and $\mathrm{var}(S_{zi}) = \sigma_{Sz}^2$, and therefore $\mathrm{var}(T_{zi}) \equiv \sigma_{Tz}^2 = \beta_z^2 \sigma_{Sz}^2 + \sigma_z^2$. To keep the presentation consistent with Section 2, we slightly reformulate the derivation in Begg and Leung (2000). Under the aforementioned linear model the observed intervention effect is

$$\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}} = \alpha_1 + \beta_1 \bar{s}_1 + \bar{\varepsilon}_1 - (\alpha_0 + \beta_0 \bar{s}_0 + \bar{\varepsilon}_0), \tag{14}$$

where the superscript 'L' indicates a linear model. Essentially, Begg and Leung (2000) defined the predicted intervention effect as

$$\begin{aligned} \hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}} &= E(\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}} | \bar{s}_0, \bar{s}_1) \\ &= \alpha_1 + \beta_1 \bar{s}_1 - (\alpha_0 + \beta_0 \bar{s}_0), \end{aligned} \tag{15}$$

where the superscript 'BL' indicates Begg and Leung. On the basis of equation (15) Begg and Leung (2000) argued that paradox 1 arises from the mathematical identity

$$\begin{aligned} \mathrm{var}(\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}}) &= \mathrm{var}\{E(\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}} | \bar{s}_0, \bar{s}_1)\} + E\{\mathrm{var}(\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}} | \bar{s}_0, \bar{s}_1)\} \\ &= \mathrm{var}(\hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}}) + E\{\mathrm{var}(\hat{\Delta}_{\mathrm{obs}}^{\mathrm{L}} | \bar{s}_0, \bar{s}_1)\} \\ &\geqslant \mathrm{var}(\hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}}). \end{aligned} \tag{16}$$

Let $\rho_{(T,S)z}$ denote the correlation between S and T in group z of the validation trial. From equation (15), $\mathrm{var}(\hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}}) = \Sigma_z \beta_z^2 \sigma_{Sz}^2$. Substituting the identity $\beta_z^2 = \rho_{(T,S)z}^2 \sigma_{Tz}^2 / \sigma_{Sz}^2$ into this latter formula gives

$$\mathrm{var}(\hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}}) = \sum_z \frac{\sigma_{Tz}^2}{N} \rho_{(T,S)z}^2. \tag{17}$$

From equation (17) Begg and Leung (2000) noted a second paradox, related to paradox 2, that, the greater the correlation between surrogate and true end points, the greater the variance of the predicted true end point.

Importantly expressions (16) and (17) do *not* assume the Prentice criterion. Begg and Leung (2000) subsequently introduced the Prentice criterion into the linear model 'to provide relatively simple insights into these counter-intuitive results'. To incorporate the Prentice criterion they essentially assumed that $\beta_1 = \beta_0$ in equation (15). By standardizing the variance, $\sigma_{Tz}^2 = \sigma_{Sz}^2 = 1$, and assuming the same correlation, $\rho_{(T,S)} = \rho_{(T,S)z}$, they obtained $E(\hat{\Delta}_{\mathrm{pred}}^{\mathrm{BL}}) = \rho_{(T,S)}\{E(\bar{s}_1) - E(\bar{s}_0)\}$. This led to further paradoxes.

Thus, although the basic paradoxes in Begg and Leung (2000) are similar to those in Day and Duffy (1996), they arise for a different reason that does not involve the Prentice criterion, namely defining the predicted intervention effect in equation (15) as a function of unknown parameters and not their estimates.

3.2. Relevant derivation with parameter estimates

To confirm our beliefs about the cause of the Begg and Leung paradoxes, we rederived the variances by substituting parameter estimates for parameters. If our explanation for the paradoxes is correct, the paradoxes should disappear. Note that we do not need to assume the Prentice criterion. We started by replacing equation (15) by

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}} = \hat{\alpha}^*_1 + \hat{\beta}^*_1 \bar{s}_1 - (\hat{\alpha}^*_0 + \hat{\beta}^*_0 \bar{s}_0), \tag{18}$$

where α^*_z and β^*_z are parameters in the previous trial and $\hat{\alpha}^*_z$ and $\hat{\beta}^*_z$ are the estimates. Let N_z and N^*_z denote the size in group z in the new and previous trials respectively. Also let $\mathrm{var}(s^*_{zi}) = \sigma^{*2}_{Sz}$, and let $\rho^*_{(T,S)z}$ denote the correlation of S and T in the previous study. As derived in Appendix A, the variance of equation (18) is

$$\mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}) = \sum_z \rho^2_{(T,S)z} \frac{\sigma^2_{Tz}}{N_z} + \sum_z (1 - \rho^{*2}_{(T,S)z}) \sigma^{*2}_{Tz} \left\{ \frac{\sigma^2_{Sz}}{\sigma^{*2}_{Sz} N_z N^*_z} + \frac{(\mu_{Sz} - \mu^*_{Sz})^2 + \sigma^{*2}_{Sz}}{\sigma^{*2}_{Sz} N^*_z} \right\}, \tag{19}$$

where $\mu_{Sz} = E(s_{zi})$ and $\mu^*_{Sz} = E(s^*_{zi})$. The first term in equation (19) is equation (17), which is the variance of the estimate from Begg and Leung (2000); the second term in equation (19) is the component of variance arising from the parameter estimates.

We check that equation (19) agrees with intuition. Suppose that in previous and validation trials the joint normal distributions of the data are the same, i.e. $\mu_z = \mu^*_z$, $\sigma^2_{Sz} = \sigma^{*2}_{Sz}$ and $\rho_{(T,S)z} = \rho^*_{(T,Sz)}$. Then equation (19) reduces to

$$\mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}) = \sum_z \frac{\sigma^2_{Tz}}{N_z} \rho^2_{(T,S)z} + \sum_z \frac{\sigma^2_{Tz}}{N_z} \frac{N_z + 1}{N^*_z} (1 - \rho^2_{(T,S)z}). \tag{20}$$

If the previous and validation trial have essentially the same size ($N^*_z = N_z + 1$), $\mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}) = \mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{obs}}) = \Sigma_z \sigma^2_{Tz}/N_z$, as expected. Intuitively, the critical sample size is $N^*_z = N_z + 1$ rather than $N^*_z = N_z$ because $\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}$ involves two parameters per group, α_z and β_z, whereas $\hat{\Delta}^{\mathrm{L}}_{\mathrm{obs}}$ requires only one, the mean. If the previous trial were larger than the validation trial ($N^*_z > N_z + 1$), then, regardless of $\rho_{(T,S)z}$, $\mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}) \leqslant \mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{obs}})$. Thus, if the variance of the estimated parameters is incorporated in the calculations of Begg and Leung (2000), the paradoxes in Begg and Leung (2000) disappear.

4. Conclusion

We found that the paradoxes in Day and Duffy (1996) arise by comparing an estimate using the surrogate end points that assumes the Prentice criterion with an estimate using the true end point that does *not* assume the Prentice criterion. If both estimates either assume or do not assume the Prentice criterion there is no paradox. We also found that the paradoxes in Begg and Leung (2000), although apparently similar to those in Day and Duffy (1996), arise for a different reason that does not involve the Prentice criterion, namely not accounting for the variability in parameter estimates. When we account for the variability in the parameter estimates, the paradoxes in Begg and Leung (2000) disappear. We conclude that there are no problems with the inferential foundations in Day and Duffy (1996), and meta-analytic extensions of the approach of Day and Duffy (1996) can be developed with confidence.

We caution that, regardless of the model, there are two inherent limitations in evaluating interventions by using surrogate end points. First, there is an inherent uncertainty about whether or not parameters that are estimated from previous trials apply to an application trial. Second, a surrogate end point for benefit does not provide information about possibly harmful side-effects

that could occur *after* the surrogate end point has been observed and before the true end point has been observed (Baker and Kramer, 2003). Thus there will always be some caution in the use of surrogate end points to evaluate interventions. Nevertheless, there are some situations, such as preliminary studies or studies to refine recommendations for interventions, where the benefits of using a surrogate end point would probably outweigh these cautions.

Acknowledgements

We thank Vance Berger, Constantine Frangakis, Mitchell Gail, Ping Hu, Ruth Pfeiffer, Don Rubin and Jian-Lun Xu for helpful discussions concerning surrogate end points.

Appendix A

For computing the variance of $\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}$, it is convenient to introduce matrix notation. Let $\theta^*_z = (\alpha^*_z, \beta^*_z)$. Also let $\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}} = \bar{S}_1\hat{\theta}^*_1 - \bar{S}_0\hat{\theta}^*_0$, where $\bar{S}_z = (1, \bar{s}_z)$ and $\hat{\theta}^*_z = (\hat{\alpha}^*_z, \hat{\beta}^*_z)$. Let S^*_z denote a matrix of data from the previous trial with rows $(1, s^*_{zi})$ and let t^*_z denote a column vector with elements t^*_{zi}. The least squares estimate of θ^*_z is $\hat{\theta}^*_z = (S^{*\mathrm{T}}_z S^*_z)^{-1} S^{*\mathrm{T}}_z t^*_z$ with variance $\mathrm{var}(\hat{\theta}^*_z) = \sigma^2_{\varepsilon z}(S^{*\mathrm{T}}_z S^*_z)^{-1}$. Let N and N^* denote the size in each group in the validation and previous trials respectively (with, for simplicity, the same sample size in each group). Define $E(s_{zi}) = \mu_{Sz}$, $E(s^*_{zi}) = \mu^*_{Sz}$, $\mathrm{var}(s_{zi}) = \sigma^2_{Sz}$ and $\mathrm{var}(s^*_{zi}) = \sigma^{*2}_{Sz}$, and let $\rho_{(T,S)z}$ and $\rho^*_{(T,S)z}$ denote the correlations in the validation and previous studies respectively. The variance of $\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}$ is

$$\mathrm{var}(\hat{\Delta}^{\mathrm{L}}_{\mathrm{pred}}) = \mathrm{var}(\bar{S}_1\hat{\theta}^*_1) + \mathrm{var}(\bar{S}_0\hat{\theta}^*_0),$$

where

$$\begin{aligned}
\mathrm{var}(\bar{S}_z\hat{\theta}^*_z) &= \mathrm{var}_{\bar{S}_z}\{E_{\hat{\theta}^*_z}(\bar{S}_z\hat{\theta}^*_z|\bar{S}_z)\} + E_{\bar{S}_z}\{\mathrm{var}_{\hat{\theta}^*_z}(\bar{S}_z\hat{\theta}^*_z|\bar{S}_z)\} \\
&= \mathrm{var}(\bar{S}_z\theta_z) + E\{\bar{S}^{\mathrm{T}}_z \mathrm{var}(\hat{\theta}^*_z)\bar{S}_z\} \\
&= \theta^{\mathrm{T}}_z \mathrm{var}(\bar{S}_z)\theta_z + \sigma^{*2}_{\varepsilon z} E[\bar{S}^{\mathrm{T}}_z\{E(S^{*\mathrm{T}}_z S^*_z)\}^{-1}\bar{S}_z] \\
&= \theta^{\mathrm{T}}_z \mathrm{var}(\bar{S}_z)\theta_z + \sigma^2_{\varepsilon z} E(\bar{S}^{\mathrm{T}}_z A^*_z \bar{S}_z),
\end{aligned} \tag{21}$$

where $A^*_z = E(S^{*\mathrm{T}}_z S^*_z)^{-1}$. The first term in equation (21) is

$$\theta^{\mathrm{T}}\begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{Sz}/N_z \end{pmatrix}\theta = \beta^2_z\sigma^2_{Sz}/N_z. \tag{22}$$

To compute the second term in equation (21), we use the result that $E(Y^{\mathrm{T}}QY) = \mathrm{tr}\{\mathrm{var}(Y)Q\} + E(Y)Q\,E(Y)^{\mathrm{T}}$, which gives

$$\sigma^2_{\varepsilon z} E(\bar{S}^{\mathrm{T}}_z A^*_z \bar{S}_z) = \sigma^2_{\varepsilon z}[\mathrm{tr}\{\mathrm{var}(\bar{S}_z)A^*_z\} + E(\bar{S}^{\mathrm{T}}_z)A^*_z\,E(\bar{S}^{\mathrm{T}}_z)] \tag{23}$$

To simplify equation (23), we write

$$\begin{aligned}
A^*_z &= E\begin{pmatrix} N^* & \sum_i s^*_{zi} \\ \sum_i s^*_{zi} & \sum_i s^{*2}_{zi} \end{pmatrix}^{-1} \\
&= \frac{1}{N^*_z}\begin{pmatrix} \sigma^{*2}_{Sz} + \mu^{*2}_{Sz} & -\mu^*_{Sz} \\ -\mu^*_{Sz} & 1 \end{pmatrix}\frac{1}{\sigma^2_{Sz}},
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{tr}\{\mathrm{var}(\bar{S}_z)A^*_z\} &= \mathrm{tr}\left\{\begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{Sz}/N_z \end{pmatrix}\frac{1}{N^*_z}\begin{pmatrix} \sigma^{*2}_{Sz} + \mu^{*2}_{Sz} & -\mu^*_{Sz} \\ -\mu^*_{Sz} & 1 \end{pmatrix}\frac{1}{\sigma^2_{Sz}}\right\} \\
&= \frac{\sigma^2_{Sz}}{\sigma^{*2}_{Sz} N_z N^*_z},
\end{aligned} \tag{24}$$

$$E(\bar{S}_z^{\mathrm{T}})A_z^* E(\bar{S}_z^{\mathrm{T}}) = (1, \mu_{Sz}) \frac{1}{N_z^*} \begin{pmatrix} \sigma_{Sz}^{*2} + \mu_{Sz}^{*2} & -\mu_{Sz}^* \\ -\mu_{Sz}^* & 1 \end{pmatrix} \frac{1}{\sigma_{Sz}^2} \begin{pmatrix} 1 \\ \mu_{Sz} \end{pmatrix}$$
$$= \frac{(\mu_{Sz} - \mu_{Sz}^*)^2 + \sigma_{Sz}^{*2}}{\sigma_{Sz}^{*2} N_z^*}. \tag{25}$$

Substituting equations (24) and (25) into equation (23) and then into equation (21) gives

$$\mathrm{var}(\bar{S}_z \hat{\theta}_z^*) = \frac{\beta_z^2 \sigma_{Sz}^2}{N_z} + \sigma_{\varepsilon z}^{*2} \left\{ \frac{\sigma_{Sz}^2}{\sigma_{Sz}^{*2} N_z N_z^*} + \frac{(\mu_z - \mu_z^*)^2 + \sigma_{Sz}^{*2}}{\sigma_{Sz}^{*2} N_z^*} \right\}. \tag{26}$$

Because $\beta_z^2 = \rho_{(T,S)z}^2 \sigma_{Tz}^2 / \sigma_{Sz}^2$ and $\sigma_{Tz}^{*2} = \beta_z^{*2} \sigma_{Sz}^2 + \sigma_{\varepsilon z}^2 = \rho_{(T,S)z}^{*2} \sigma_{Tz}^{*2} + \sigma_{\varepsilon z}^{*2}$, we have $\sigma_{\varepsilon z}^{*2} = (1 - \rho_{(T,S)z}^{*2}) \sigma_{Tz}^{*2}$, and we can write equation (26) as

$$\mathrm{var}(\bar{S}_z \hat{\theta}_z^*) = \rho_{(T,S)z}^2 \frac{\sigma_{Tz}^2}{N_z} + (1 - \rho_{(T,Sz)}^{*2}) \sigma_{Tz}^{*2} \left\{ \frac{\sigma_{Sz}^2}{\sigma_{Sz}^{*2} N_z N_z^*} + \frac{(\mu_z - \mu_z^*)^2 + \sigma_{Sz}^{*2}}{\sigma_{Sz}^{*2} N_z^*} \right\}. \tag{27}$$

This leads to equation (19) in the text.

References

Baker, S. G. and Kramer, B. S. (2003) A perfect correlate does not a surrogate make. *BMC Med. Res. Methodol.*, **3**, 16.

Begg, C. B. and Leung, D. H. Y. (2000) On the use of surrogate end points in randomized trials (with comments). *J. R. Statist. Soc.* A, **163**, 15–28.

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

Day, N. E. and Duffy, S. W. (1996) Trial design based on surrogate end points—application to comparison of different breast screening frequencies. *J. R. Statist. Soc.* A, **159**, 49–60.

Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992) Statistical validation of intermediate endpoints for chronic disease. *Statist. Med.*, **11**, 167–178.

Gail, M. H., Pfeiffer, R., Houwelingen, H. C. and Carroll, R. J. (2000) On meta-analytic assessment of surrogate outcomes. *Biostatistics*, **3**, 231–246.

Morrison, A. S. (1991) Intermediate determinants of mortality in the evaluation of screening. *Int. J. Epidem.*, **20**, 642–650.

Prentice, R. L. (1989) Surrogate end points in clinical trials: definitions and operational criteria. *Statist. Med.*, **8**, 431–440.