Variance and sample size reduction using surrogate end-points

Technical Report No.ASD/2012/12
Dated: 12 December 2012

Buddhananda Banerjee
And
Prof. Atanu Biswas

Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108

buddhananda_r@isical.ac.in, atanu@isical.ac.in

Variance and sample size reduction using surrogate end-points

Buddhananda Banerjee and Atanu Biswas
Applied Statistics Unit, Indian Statistical Institute
203 B. T. Road, Kolkata – 700 108, India
buddhananda_r@isical.ac.in, atanu@isical.ac.in

Abstract: Surrogate end-points are used when the true end-points are costly or time-consuming. In a typical set up we observe a fixed proportion of true-and-surrogate responses, and the remaining proportion are only-surrogate responses. It is obvious that the inclusion of such only-surrogate end-points increase the efficiency of associated estimation. In this present paper we want to quantify the (inverse of) efficiency as a function of the proportion of available true responses. Also we obtain the expression of the gain in true sample size at the expense of surrogates to achieve a fixed power, as a function of the proportion of true responses. We present our discussion in the two-treatment set up in the context of odds ratio. We illustrate the procedure using some real data set.

Keywords: Surrogate responses; Odds ratio; Risk ratio; Treatment difference.

1 Introduction

Surrogate end-point is chosen as a measure or indicator of a biological process that is obtained sooner, at less cost than a true end-point of health outcome, and is used to arrive at conclusions about the effect of intervention on the true end-point. Surrogate end-points are used with growing interest in medical science. For example, in a trial of treatment for osteoporosis we might be interested in reducing the fracture rate, but we measure bone mineral density (BMD) instead. A change in CD4 cell count in randomized trial is considered as a surrogate for survival time in the study of HIV. Again, damage to the heart muscle due to myocardial infraction can be accurately assessed by arterioscintography. As it is an expensive procedure, peak cardiac enzyme level in the blood stream, which is more easily obtainable, is used as surrogate measure

of heart vascular damage (see [1]). Sometimes observed value of response variable in the middle of an experiment is considered as the surrogate end-point. For example patients with age related macular degeneration (ARMD) progressively loose vision. To compare between placebo and high-dose interferon-α for its treatment observations are taken after six months and one year. The observation at six-month is taken as a surrogate of the final one-year response.

Prentice [2] gave validation criteria for a surrogate, which is further discussed by Freedman, Graubard and Schatzkin [3], Reilly and Pepe [4], Day and Duffy [5], Buyse and Molenberghs [6], Buyse et al. [7], Molenberghs, Geys and Buyse [8], and Chen, Geng and Jia [9], among others. The use of surrogate end-points is likely to be beneficial, not only in time or monitory sense, but it gives more accuracy in estimation of target parametric functions such as treatment difference, odds ratio, etc. Pepe [10] obtained the distribution of estimator of regression parameter when the validation sample fraction has a fixed limiting value, ρ, say. Banerjee and Biswas [11] established that the variance of the estimator of treatment difference is bounded for fixed ρ. Lin et al. [12] and Wang and Taylor [13] defined the proportion of treatment effect explained by the surrogate for hazard model. Chen [14] and Begg and Leung [15] discussed the inferential improvement by the use of surrogate end-points. Chen et al. [16] introduced the concept of information recovery from surrogate end-points by considering linear models for true and surrogate on covariates.

Proportion of validation sample, ρ, naturally plays a key role in the gain in associated inference. The validation samples are true and surrogate paired observation, but the rest of the samples are surrogate responses only. In Section 2 we describe the set up in details and the data structure under a general probability model for binary true and and binary surrogate response. In Section 3 we establish that the (inverse of) relative efficiency to estimate the treatment success probability by using surrogate end-points is a linear function of the validation sample proportion, ρ. As a simple consequence of that we also prove the (inverse of) relative efficiency to estimate treatments difference and log odds ratio in a two-treatment set up is also linear in ρ_A and ρ_B, the validation sample proportions for the two treatments A and B, respectively. In Section 4 establish

that the sample size reduction with respect to the true end-points (to attain the same power at same level for a test of treatment equivalence based on odds ratio) is also a linear function of ρ_A and ρ_B. An illustration with a real data set is given in Section 5. Section 6 concludes.

2 Experiment detail and data structure

We consider a set up of two treatment binary end-points with binary surrogates. Begg and Leung [15] pointed out that for binary end-points, the probability of concordance is an indicator of association between true and surrogate end-points. Suppose n_A and n_B patients are allotted to treatments A and B, respectively; but we get only m_A and m_B true end-points along with all surrogate end-points within the stipulated time frame or cost limit, where $m_i << n_i$, $i = A, B$. Let Z be an indicator variable such that $Z = 1$ or 0 according as the treatment A or B is allocated to a patient. Denote true and surrogate end-points for treatment t by Y_t and W_t, $t = A, B$. All these end-points are either 1 or 0 for success or failure, respectively. We denote $p_t = P(Y_t = 1)$ as the success probability by the true end-points for treatment t. Furthermore, let us denote

$$P(W_t = 1|Y_t = 1) = \pi_{t1} \quad \text{and} \quad P(W_t = 0|Y_t = 0) = \pi_{t0}, \tag{2.1}$$

which are the *sensitivity* and *specificity* of the 2×2 table for treatment A where the true and surrogate responses are in the two margins. Consequently, the success probabilities by the surrogate responses for the two treatments are,

$$r_t = P(W_t = 1) = (1 - \pi_{t0}) + (\pi_{t1} + \pi_{t0} - 1)p_t.$$

The data corresponding to treatment t can be represented in a table as follows.

True↓ Surrogate→	$W_t = 1$	$W_t = 0$	Total
$Y_t = 1$	m_{t11}	m_{t10}	Y_{tT}
$Y_t = 0$	m_{t01}	m_{t00}	$m_t - Y_{tT}$
Total	W_{tT}	$m_t - W_{tT}$	m_t

where $Y_{tT} = \sum_{i=1}^{m_t} Y_{ti}$ and $W_{tT} = \sum_{i=1}^{m_t} W_{ti}$; also we denote $W_{tS} = \sum_{i=m_t+1}^{n_t} W_{ti}$. If any marginal is found to be zero, it is customary to add 0.5 to each of the marginals. Banerjee and Biswas [11] considered a similar set set.

3 Variance reduction using surrogate responses

Consider the likelihood for treatment t,

$$\begin{aligned} L(\xi_t) &= \binom{m_t}{y_{tT}} p_t^{y_{tT}} q_t^{m_t - y_{tT}} \binom{y_{tT}}{m_{t11}} \pi_{t1}^{m_{t11}} (1-\pi_{t1})^{y_{tT}-m_{t11}} \\ &\quad \times \binom{m_t - y_{tT}}{m_{t01}} (1-\pi_{t0})^{m_{t01}} \pi_{t0}^{m_{t00}} \binom{n_t - m_t}{w_{tS}} r_t^{w_{tS}} (1-r_t)^{n_t - m_t - w_{tS}}, \quad (3.1) \end{aligned}$$

where $\xi_t = (p_t, \pi_{t1}, \pi_{t0})$ and $q_t = 1 - p_t$. The Fisher information matrix is $\mathbf{I}(\xi_t)$ and the $(1,1)$th element of $[\mathbf{I}(\xi_t)]^{-1}$, denoted by $[\mathbf{I}(\xi_t)]_{11}^{-1}$, gives the variance of $\widehat{p}_t$. Here we define the inverse of efficiency, which is the measure of improvement in estimation of p_t by surrogate augmented analysis as $\frac{[\mathbf{I}(\xi_t)]_{11}^{-1}}{p_t q_t / m_t}$.

Furthermore denote $m_t/n_t = \rho_t \in (0,1]$. But $\rho_t = 0$ only when $m_t = 0$, indicating no true response is available. This is not of any statistical interest. Using $m_t = \rho_t n_t$ we get $\mathbf{I}(\xi_t) = n_t \boldsymbol{I}_{\xi_t}(\rho_t)$. Hence the measure of improvement, given by the proportional variance, reduces to

$$G_{\xi_t}(\rho_t) = \frac{n_t^{-1} \left[\boldsymbol{I}_{\xi_t}(\rho_t)\right]_{11}^{-1}}{m_t^{-1} p_t q_t} = \frac{\rho_t \left[\boldsymbol{I}_{\xi_t}(\rho_t)\right]_{11}^{-1}}{p_t q_t}.$$

Now we have the following theorems whose proofs are given in the appendix.

Theorem 1: *(a) Relative gain, denoted by the inverse of efficiency, $G_{\xi_t}(\rho_t)$ by using surrogate end-points with ρ_t proportion of available true responses is a linear function of ρ_t, that is*

$$G_{\xi_t}(\rho_t) = C_t + (1 - C_t)\rho_t = \rho_t + (1 - \rho_t)C_t, \quad (3.2)$$

which is a straight line joining the points $(0, C_t)$ and $(1,1)$ with intercept and slope adds to unity, with

$$C_t = \frac{q_t \pi_{t0}(1-\pi_{t0}) + p_t \pi_{t1}(1-\pi_{t1})}{r_t(1-r_t)}.$$

(b) Further

$$C_t = \frac{E(Var(W_t|Y_t))}{Var(W_t)} \in [0,1].$$

Theorem 2: *The inverse of efficiency in variance estimation by using surrogate end-points to estimate the log odds ratio (OR),* $\theta = \log\left(\frac{p_A q_B}{q_A p_B}\right)$*, is given by a plane, that is*

$$G_{OR}(\rho_A, \rho_B) = \frac{(m_A p_A q_A)^{-1} G_{\xi_A}(\rho_A) + (m_B p_B q_B)^{-1} G_{\xi_B}(\rho_B)}{(m_A p_A q_A)^{-1} + (m_B p_B q_B)^{-1}} \tag{3.3}$$

Corollary 1: When $\rho_A = \rho_B = \rho$, the inverse of efficiency by using surrogate end-points to estimate the log odds ratio is a linear function of ρ, that is

$$G_{OR}(\rho) = \rho + (1-\rho)\left\{\frac{(n_A p_A q_A)^{-1} C_A + (n_B p_B q_B)^{-1} C_B}{(n_A p_A q_A)^{-1} + (n_B p_B q_B)^{-1}}\right\}.$$

Corollary 2: The inverse of efficiency by using surrogate end-points to estimate $\theta = p_A - p_B$, the treatment difference (TD), is a plane given by

$$G_{TD}(\rho_A, \rho_B) = \frac{m_A^{-1} p_A q_A G_{\xi_A}(\rho_A) + m_B^{-1} p_B q_B G_{\xi_B}(\rho_B)}{m_A^{-1} p_A q_A + m_B^{-1} p_B q_B},$$

and, for $\rho_A = \rho_B = \rho$ we get the line given by

$$G_{TD}(\rho) = \rho + (1-\rho)\left\{\frac{n_A^{-1} p_A q_A C_A + n_B^{-1} p_B q_B C_B}{n_A^{-1} p_A q_A + n_B^{-1} p_B q_B}\right\}.$$

Corollary 3: The inverse of efficiency by using surrogate end-points to estimate $\theta = \log\left(\frac{p_A}{p_B}\right)$, the log risk ratio (RR), is a plane given by

$$G_{RR}(\rho_A, \rho_B) = \frac{(m_A p_A)^{-1} q_A G_{\xi_A}(\rho_A) + (m_B p_B)^{-1} q_B G_{\xi_B}(\rho_B)}{(m_A p_A)^{-1} q_A + (m_B p_B)^{-1} q_B},$$

and when $\rho_A = \rho_B = \rho$ it reduces to a linear function of ρ given by

$$G_{RR}(\rho) = \rho + (1-\rho)\left\{\frac{(n_A p_A)^{-1} q_A C_A + (n_B p_B)^{-1} q_B C_B}{(n_A p_A)^{-1} q_A + (n_B p_B)^{-1} q_B}\right\}.$$

In Figure 1 we plot $G_{OR}(\rho_A, \rho_B)$, given in 3.3, against ρ_A and ρ_B. We find that, for suitable ρ_A and ρ_B, there is considerable gain in estimation of OR. Figure 2 illustrates the special case of $\rho_A = \rho_B$ Here G_A, G_B, G_{OR}, G_{TD}, G_{RR} are all straight lines. In fact, G_{OR}, G_{TD} and G_{RR} are convex combinations of G_A and G_B, and belong to in between these two straight lines. We have taken $m_A = m_B = 34$ and $p_A = 0.7$, $p_B = 0.8$ in these two figures, for illustration.

4 Sample size reduction

Now our objective is to assess the gain in true sample size by using surrogate end-points, assuming that the true end-points are difficult and/or costly to get. Let us consider log odds ratio only for illustration.

Consider the testing problem to compare the treatment effects in terms of log odds ratio when surrogate data are present, that is for the odds ratio $\psi = \frac{p_A q_B}{q_A p_B}$, we want to test $H_0 : \log \psi = 0$ against one sided or both sided alternative. First we find the sample size required to achieve $(1 - \beta)$ power in a level α test, and then obtain the reduction in sample size when surrogate responses are present.

Here for simplicity we first consider the special case of $\rho_A = \rho = \rho_B$, and assume that $m_A/m_B = k$ is pre-specified. We first list the notations that we intend to use.

m_t^T : number of true end-points required for treatment t when only true end-points are used;

n_t^S : total number of surrogate end-points required for treatment t when ρ_t proportion of true end-points are used;

m_t^S : number of true end-points required for treatment t when ρ_t proportion true end-points are used, i.e. $m_t^S = \rho_t n_t^S$.

Based on only available true end-points $T_1 = \log\left(\frac{Y_{AT}(m_B - Y_{BT})}{(m_A - Y_{AT})Y_{BT}}\right)$ is an estimator of $\log \psi$. It is well-known that T_1 is an asymptotically unbiased for $\log \psi$, and also asymptotically normal with variance given by (7.2). If the parameters are assumed to be known then for a both sided test, the required true sample size for treatment B will be

$$m_B^T = \left(\frac{z_{\alpha/2} + z_\beta}{\log \psi}\right)^2 \left\{\frac{1}{k p_A q_A} + \frac{1}{p_B q_B}\right\},$$

where z_γ is the upper γ percentile of the standard normal distribution, and for treatment A it will be $m_A^T = k.m_B^T$. When the parameters are not known, their estimated values are to be used. If surrogate responses are present, $\log \psi$ can be estimated by $T_2 = \log\left(\frac{\widehat{\mathbf{Y}}_A(n_B - \widehat{\mathbf{Y}}_B)}{(n_A - \widehat{\mathbf{Y}}_A)\widehat{\mathbf{Y}}_B}\right)$ whose asymptotic variance is given by (7.3). Thus we immediately get the total surrogate sample size for treatment B as a function of parameters

as

$$n_B^S = \left(\frac{z_{\alpha/2} + z_\beta}{\log \psi}\right)^2 \left\{ \frac{[\boldsymbol{I}_{\xi_A}(\rho_A)]_{11}^{-1}}{k(p_A q_A)^2} + \frac{[\boldsymbol{I}_{\xi_B}(\rho_B)]_{11}^{-1}}{(p_B q_B)^2} \right\}.$$

The proportion of true samples for treatment B for the surrogate-augmented situation to the no-surrogate situation is a plane

$$\frac{m_B^S}{m_B^T} = \frac{\rho_B n_B^S}{m_B^T} = \left\{ \frac{(kp_A q_A)^{-1} G_{\xi_A}(\rho_A) + (p_B q_B)^{-1} G_{\xi_B}(\rho_B)}{(kp_A q_A)^{-1} + (p_B q_B)^{-1}} \right\} = G(\rho_A, \rho_B).$$

So $m_A^S = k.m_B^S = k.G(\rho_A, \rho_B) m_B^T$ which implies that total number of true end-points reduction is $m_A^S + m_B^S = (1+k)G(\rho_A, \rho_B) m_B^T$. Proportion of total true end-points reduction is

$$\frac{m_A^S + m_B^S}{m_A^T + m_B^T} = \frac{(1+k)G(\rho_A, \rho_B) m_B^T}{(1+k) m_B^T} = G(\rho_A, \rho_B), \tag{4.1}$$

which is shown in Figure 3 for $k = 1$ and $p_A = 0.7$, $p_B = 0.8$. It is observed that with sufficient small values of ρ_A and ρ_B we the reduction in true end-points is remarkable.

Reduction of true end-point from for treatment B from m_B^T to $G(\rho_A, \rho_B) m_B^T$ will be compensated by an additional $\frac{1-\rho_B}{\rho_B} G(\rho_A, \rho_B) m_B^T$ only-surrogate end-points, and for treatment A the number will be $\frac{1-\rho_A}{\rho_A} G(\rho_A, \rho_B) k m_B^T$. So total number of only surrogate end-points required is

$$G(\rho_A, \rho_B) \left\{ \frac{(1-\rho_A)k}{\rho_A} + \frac{(1-\rho_B)}{\rho_B} \right\} m_B^T \tag{4.2}$$

which is

$$G^S(\rho_A, \rho_B) = \frac{G(\rho_A, \rho_B)}{1+k} \left\{ \frac{(1-\rho_A)k}{\rho_A} + \frac{(1-\rho_B)}{\rho_B} \right\} \tag{4.3}$$

times required number of true end-points when only true end-points are used (i.e. $(1+k)m_B^T$). The equation (4.3) is presented by the dark shaded surface in Figure 4 for $p_A = 0.7$, $p_B = 0.8$ again.

When $\rho_A = \rho_B = \rho$, it simplifies to

$$\frac{m_B^S}{m_B^T} = \frac{\left\{ \frac{C_A}{kp_A q_A} + \frac{C_B}{p_B q_B} \right\}}{\left\{ \frac{1}{kp_A q_A} + \frac{1}{p_B q_B} \right\}} + \frac{\left\{ \frac{1-C_A}{kp_A q_A} + \frac{1-C_B}{p_B q_B} \right\}}{\left\{ \frac{1}{kp_A q_A} + \frac{1}{p_B q_B} \right\}} \rho = G(\rho).$$

So total number of only surrogate end-points required is $\frac{1-\rho}{\rho} G(\rho)(1+k) m_B^T$ which is $\frac{1-\rho}{\rho} G(\rho)$ times required number of true end-points when only true end-points are used.

5 Illustration with real data example

Consider the data set used by Chuang [17] to study categorical covariate using monotone scores model. A group of patients suffering from chronic insomnia were randomly assigned to receive either an active CNS drug (A: $Z = 1$) or a placebo (B: $Z = 0$) in a single-blind study. Treatment procedures are evaluated after the first week (surrogate, W) and the second week (true response, Y) for its helpfulness. Each patient was asked whether the medication helped him/her, and if yes, then how much. Denote two categories in end-points namely 'Got Help' (where medication helped, either 'a lot' or 'quite a bit' or 'a little') and 'No Help' for both the surrogate and true responses. The data set is analysed by Banerjee and Biswas (2011). The estimated $p_A = 0.8095$, $p_B = 0.4882$, $\pi_{1A} = 0.3039$, $\pi_{0A} = 0.7917$, $\pi_{1B} = 0.3710$, $\pi_{0B} = 0.1077$. Treating these as the true values, in Table 1 we provide $m_A^S + m_B^S = G(\rho_A, \rho_B)(m_A^T + m_B^T)$, the number of true sample needed in a surrogate-augmented situation when ρ_t and $1 - \rho_t$ are the proportions of 'true-and-surrogate' and 'only surrogate' responses from treatment t, $t = A, B$, to achieve an 80% power at 5% level, see the equation (4.1). Here we treat $k = 1$ and we get $m_A^T + m_B^T = 369.7574$. Clearly the proportion of true end-points are small for small values of ρ_A and ρ_B. In Table 2 we provide the number of only-surrogate end-points given by (4.2). Here also the number increases for small values of ρ_A and ρ_B.

6 Discussions and remarks

In this article we discussed the impact of proportion of available true end-point in reduction of variance to estimate treatment success probabilities and related parametric functions like treatment difference, log odds ratio, risk ratio. We interestingly found that the proportion of variance reduction to estimate treatment success probability for individual treatment is a linear function $G_{\xi_t}(\rho_t)$ of true end-point sample proportion (ρ_t) when compared to the total (surrogate) end-points. Proportion of variance for the parametric functions are convex combinations of $G_{\xi_A}(\rho_A)$ and $G_{\xi_B}(\rho_B)$ and the weights are proportional to the variance of the estimator of the parametric functions

for individual treatments with true end-points only.

When true and surrogate end-points are independent that is $C_A = 1 = C_B$, and there is no reduction in sample size. On the contrary when the conditional distribution of surrogate end-point given the true is degenerate one i.e. $C_A = 0 = C_B$, the proportion of sample size reduction is $(1-\rho)$ when $\rho_A = \rho_B = \rho$.

For estimation we suggested to use MLE in this paper, which is iterative for the problem under consideration. For practical implementation of surrogate-augmented procedure, Banerjee and Biswas (2011) used EM-based estimates of p_A and p_B. Alternative estimates based on conditional expectations may be

$$\widehat{p}_t = \widehat{\mathbf{Y}}_t/n_t = n_t^{-1}\left\{Y_{tT} + \frac{m_{t11}}{W_{tT}}W_{tS} + \frac{m_{t10}}{m_t - W_{tT}}(n_t - m_t - W_{tS})\right\}.$$

Here the three terms within brace in the right hand side correspond to the observed number of successes from the true responses, estimate of the true successes out of W_{tS} surrogate successes for which true responses are unobserved, and estimate of true successes out of $(n_t - m_t - W_{tS})$ surrogate failures for which true responses are unobserved. Our detailed simulation studies show that the behaviour of these two estimators are almost similar to the MLE. The details are under study.

7 Appendices

7.1 Appendix 1: Proof of Theorem 1

(a) From the likelihood equation (3.1), it is immediate that

$$n_t \boldsymbol{I}_{\xi_t}(\rho_t) \equiv n_t \begin{pmatrix} \rho_t d_{t1} + (1-\rho_t)d_{t2} & (1-\rho_t)c_{t1} & (1-\rho_t)c_{t2} \\ (1-\rho_t)c_{t1} & \rho_t d_{t3} + (1-\rho_t)d_{t4} & (1-\rho_t)c_{t3} \\ (1-\rho_t)c_{t2} & (1-\rho_t)c_{t3} & \rho_t d_{t5} + (1-\rho_t)d_{t6} \end{pmatrix},$$

where $d_{t1} = \frac{1}{p_t q_t}$, $d_{t2} = \frac{(\pi_{t1}+\pi_{t0}-1)^2}{r_t(1-r_t)}$, $d_{t3} = \frac{p_t}{\pi_{t1}(1-\pi_{t1})}$, $d_{t4} = \frac{p_t^2}{r_t(1-r_t)}$, $d_{t5} = \frac{q_t}{\pi_{t0}(1-\pi_{t0})}$, $d_{t6} = \frac{q_t^2}{r_t(1-r_t)}$ and $c_{t1} = \left(\frac{\pi_{t0}}{(1-r_t)} - \frac{1-\pi_{t0}}{r_t}\right) = \frac{r_t(\pi_{t1}+\pi_{t0}-1)}{r_t(1-r_t)}$, $c_{t2} = \left(\frac{1-\pi_{t1}}{(1-r_t)} - \frac{\pi_{t1}}{r_t}\right) = -\frac{(1-r_t)(\pi_{t1}+\pi_{t0}-1)}{r_t(1-r_t)}$, $c_{t3} = -\frac{p_t q_t}{r_t(1-r_t)}$. Now observing that

$$d_{t6}d_{t4} = c_{t3}^2$$

$$
\begin{aligned}
d_{t2}d_{t4} &= c_{t1}^2 \\
d_{t6}d_{t2} &= c_{t2}^2 \\
d_{t2}d_{t4}d_{t6} + d_{t1}d_{t4}d_{t6} + d_{t2}d_{t3}d_{t6} &= d_{t1}c_{t3}^2 + d_{t3}c_{t2}^2 + d_{t5}c_{t1}^2 \qquad (7.1)\\
d_{t2}d_{t4}d_{t6} + 2c_{t1}c_{t2}c_{t3} &= d_{t2}c_{t3}^2 + d_{t4}c_{t2}^2 + d_{t6}c_{t1}^2 \\
d_{t1}d_{t3}d_{t5} &= d_{t1}d_{t4}d_{t5} + d_{t2}d_{t3}d_{t5} + d_{t1}d_{t3}d_{t6} = \frac{1}{\pi_{t1}(1-\pi_{t1})\pi_{t0}(1-\pi_{t0})},
\end{aligned}
$$

we get

$$
C_t = \frac{d_{t6}}{d_{t5}} + \frac{d_{t4}}{d_{t3}} = \frac{q_t\pi_{t0}(1-\pi_{t0}) + p_t\pi_{t1}(1-\pi_{t1})}{r_t(1-r_t)},
$$

and

$$
G_{\xi_t}(\rho_t) = \frac{n_t^{-1}\left[\boldsymbol{I}_{\xi_t}(\rho_t)\right]_{11}^{-1}}{m_t^{-1}p_tq_t} = \frac{\rho_t\left[\boldsymbol{I}_{\xi_t}(\rho_t)\right]_{11}^{-1}}{p_tq_t} = \rho_t + (1-\rho_t)C_t.
$$

For (b) we observe that $Var(W_t|Y_t = j) = \pi_{tj}(1-\pi_{tj})$ for $j = 1, 0$ and $t = A, B$. Hence

$$
E(Var(W_t|Y_t)) = q_t\pi_{t0}(1-\pi_{t0}) + p_t\pi_{t1}(1-\pi_{t1}),
$$

and hence immediately we get

$$
C_t = \frac{E(Var(W_t|Y_t))}{Var(W_t)} = \frac{E(Var(W_t|Y_t))}{E(Var(W_t|Y_t)) + Var(E(W_t|Y_t))} \in [0,1].
$$

Clearly $C_t = 1$ when true and surrogate end-points are independent, and $C_t = 0$ when the conditional distribution of surrogate end-point given the true is degenerate one, for treatment t.

7.2 Appendix 2: Proof of Theorem 2

The estimator of $\theta = \log\left(\frac{p_Aq_B}{q_Ap_B}\right)$ based on only true end-points will be

$$
\begin{aligned}
T_1 &= \log\left(\frac{Y_{AT}}{m_A - Y_{AT}}\right) - \log\left(\frac{Y_{BT}}{m_B - Y_{BT}}\right) \\
&\approx \log\left(\frac{p_A}{1-p_A}\right) + \frac{Y_{AT} - m_Ap_A}{m_Ap_A(1-p_A)} - \log\left(\frac{p_B}{1-p_B}\right) - \frac{Y_{BT} - m_Bp_B}{m_Bp_B(1-p_B)},
\end{aligned}
$$

with variance

$$
Var(T_1) \approx \frac{1}{m_Ap_Aq_A} + \frac{1}{m_Bp_Bq_B}. \qquad (7.2)
$$

But when the surrogate end-points are used along with the true end-points, the estimator will change to

$$\begin{aligned} T_2 &= \log\left(\frac{\widehat{\mathbf{Y}}_A}{n_A - \widehat{\mathbf{Y}}_A}\right) - \log\left(\frac{\widehat{\mathbf{Y}}_B}{n_B - \widehat{\mathbf{Y}}_B}\right) \\ &\approx \log\left(\frac{p_A}{1-p_A}\right) + \frac{\widehat{\mathbf{Y}}_A - n_A p_A}{n_A p_A(1-p_A)} - \log\left(\frac{p_B}{1-p_B}\right) - \frac{\widehat{\mathbf{Y}}_B - n_B p_B}{n_B p_B(1-p_B)}, \end{aligned}$$

with $\widehat{\mathbf{Y}}_t$ is the estimate of total number of successes (using the MLE's). Clearly

$$Var(T_2) \approx \frac{[\boldsymbol{I}_{\xi_A}(\rho_A)]^{-1}_{11}}{n_A(p_A q_A)^2} + \frac{[\boldsymbol{I}_{\xi_B}(\rho_B)]^{-1}_{11}}{n_B(p_B q_B)^2}. \tag{7.3}$$

Hence the relative proportion of variances (inverse of efficiency) due to the usage of surrogate end-points is

$$G_{OR}(\rho_A, \rho_B) = \frac{\frac{[\boldsymbol{I}_{\xi_A}(\rho_A)]^{-1}_{11}}{n_A(p_A q_A)^2} + \frac{[\boldsymbol{I}_{\xi_B}(\rho_B)]^{-1}_{11}}{n_B(p_B q_B)^2}}{\frac{1}{m_A p_A q_A} + \frac{1}{m_B p_B q_B}} = \frac{(m_A p_A q_A)^{-1} G_{\xi_A}(\rho_A) + (m_B p_B q_B)^{-1} G_{\xi_B}(\rho_B)}{(m_A p_A q_A)^{-1} + (m_B p_B q_B)^{-1}}.$$

Reference

1. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in Medicine* 1989; **8**: 415-425.

2. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8** : 431–440.

3. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* 1992; **11**: 167–178.

4. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995; **82**: 299–314.

5. Day NE, Duffy SW. Trial design based on surrogate endpoints - applications to comparison of different breast screening frequencies. *Journal of the Royal Statistical Society, Series A* 1996; **159**: 49–60.

6. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 1014–1029.

7. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys, H. The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biometrics* 2000;**56**: 49–67.

8. Molenberghs G, Geys H, Buyse M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine* 2001; **20**: 3023–3038.

9. Chen H, Geng Z, Jia J. Criteria for surrogate endpoints. *Journal of the Royal Statistical Society, Series B* 2007;**69**: 919–932.

10. Pepe MS. Inference using surrogate outcome data and a validation sample. *Biometrika* 1992;**79**: 355–365

11. Banerjee B, Biswas A. Estimating treatment difference for binary responses in the presence of surrogate endpoints. *Statistics in Medicine* 2011; **30**: 186–196

12. Lin DY, Fleming TR, Gruttola V De. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* 1997; **16**: 1515–1527.

13. Wang Y, Taylor JMG. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* 2002; **58**: 803–812.

14. Chen YH. A robust imputation method for surrogate outcome data. *Biometrika* 2000; **87**: 711–716.

15. Begg CB, Leung DHY. On the use of surrogate endpoints in randomized trials. *Journal of the Royal Statistical Society, Series A* 2000; **163**: 15–28.

16. Chen SX, Leung, DHY, Qin J. Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association* 2003; **98**: 1052–1062.

17. Chuang C. Analyzing an pharmaceutical data set with categorical covariate using monotone scores model.*Communications in Statistics – Computation and Simulation* 1987; **16**: 1–15.

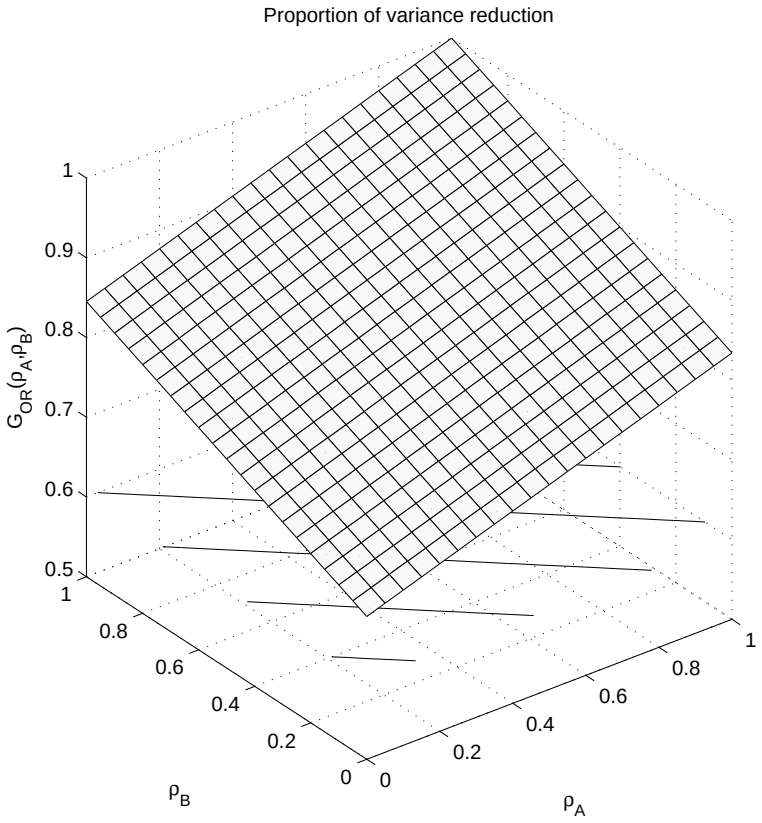

Figure 1. Proportion of variance against ρ_A and ρ_B.

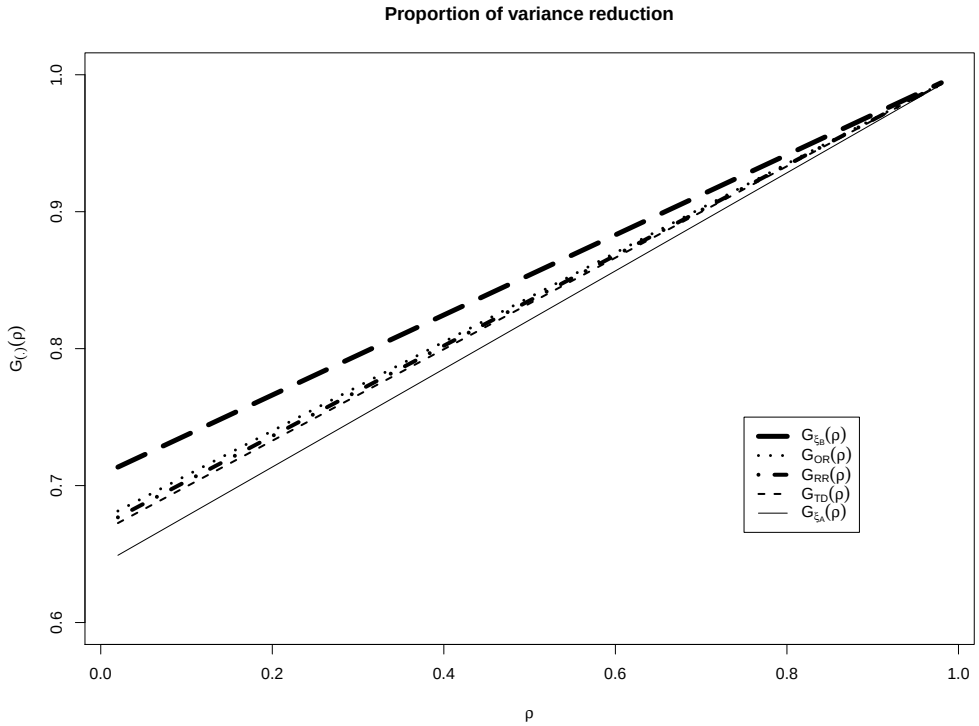

Figure 2. G_A, G_B, G_{OR}, G_{TD}, G_{RR} against ρ.

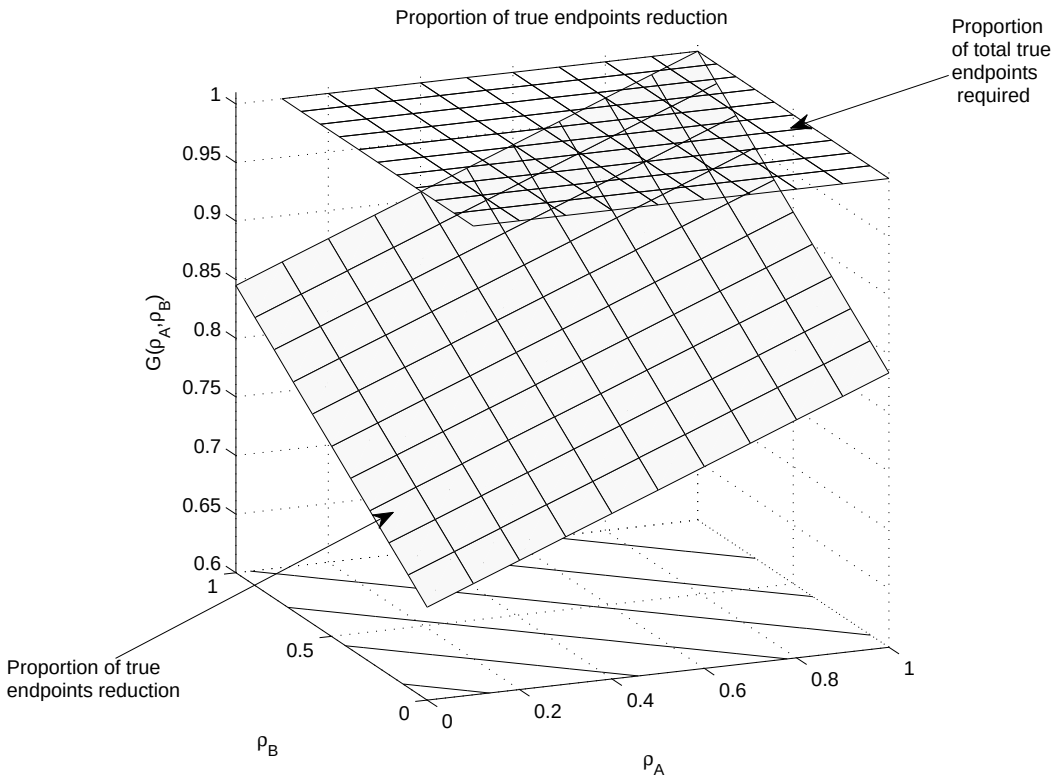

Figure 3. Proportion of total true end-points against ρ_A and ρ_B.

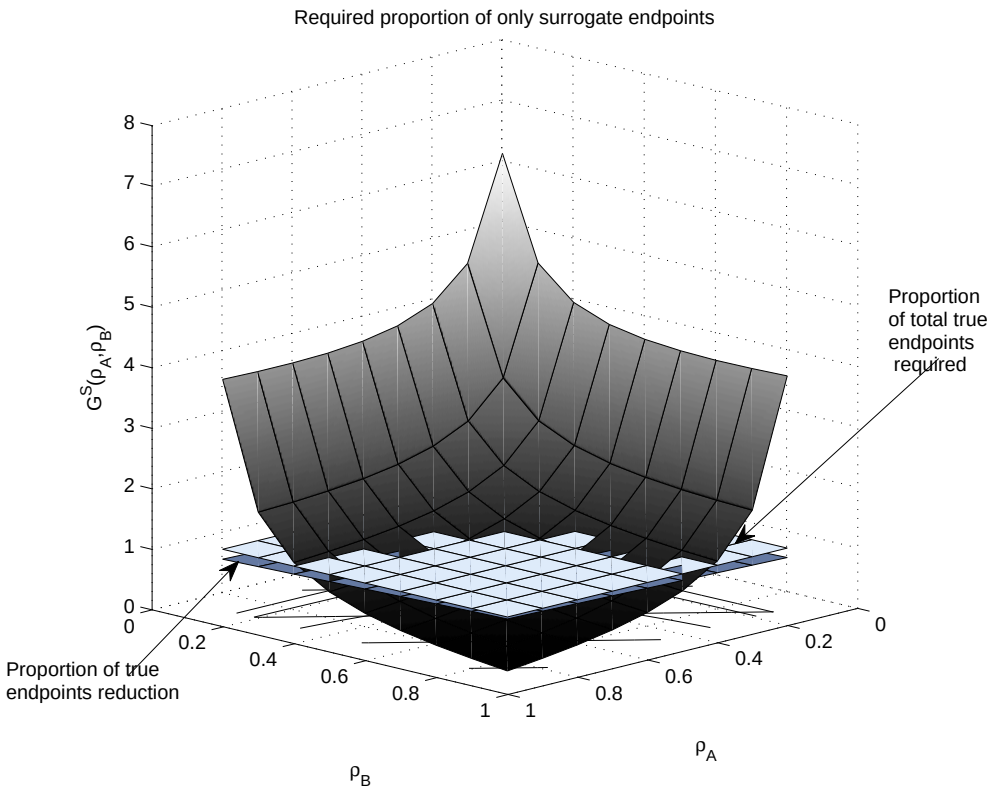

Figure 4. Proportion of only surrogate end-points required is against ρ_A and ρ_B.

Table 1: Number of true end-points required (in nearest integers) for the CNS data set up, against different values of ρ_A, ρ_B.

ρ_A \ ρ_B	0.05	0.2	0.35	0.5	0.65	0.8	0.95
0.05	280	285	289	294	298	302	307
0.2	290	294	299	303	308	312	317
0.35	300	304	309	313	317	322	326
0.5	309	314	318	323	327	332	336
0.65	319	323	328	332	337	341	346
0.8	329	333	338	342	346	351	355
0.95	338	343	347	352	356	361	365

Table 2: Number of only-surrogate end-points required (in nearest integers) for the CNS data set up, against different values of ρ_A, ρ_B.

ρ_A \ ρ_B	0.05	0.2	0.35	0.5	0.65	0.8	0.95
0.05	5326	3274	3016	2936	2912	2911	2924
0.2	3335	1178	875	758	698	663	642
0.35	3125	891	573	447	380	339	312
0.5	3094	785	455	323	252	207	177
0.65	3117	734	393	256	181	135	102
0.8	3164	708	356	214	137	88	54
0.95	3224	695	332	185	105	55	19