Statistics and Probability Letters 96 (2015) 102–108

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Linear increment in efficiency with the inclusion of surrogate endpoint

Buddhananda Banerjee [a], Atanu Biswas [b,*]

[a] Indian Institute of Science Education and Research, Kolkata, India
[b] Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 10 May 2014
Received in revised form 10 September 2014
Accepted 11 September 2014
Available online 27 September 2014

Keywords:
Surrogate responses
Odds ratio
Risk ratio
Treatment difference

ABSTRACT

In a two-sample clinical trial, a fixed proportion of true-and-surrogate and the remaining only-surrogate responses are observed. We quantify the increase in efficiency to compare the treatments as a linear function of the proportion of available true responses.

1. Introduction

A surrogate endpoint is chosen as a measure or an indicator of a biological process. Usually it is obtained sooner, at a lesser cost than the true endpoint of health outcome, and is used to arrive at a conclusion about an effect of intervention on the true endpoint. Surrogate endpoints are used with growing interest in medical science. For example, in a trial of a treatment of osteoporosis we might be interested in reduction of the fracture rate, but we measure the bone mineral density (BMD) instead. A change in CD4 cell count in a randomized trial is considered as a surrogate of survival time in the study of HIV affected patients. Again, some damages to the heart muscle due to myocardial infarction can be accurately assessed by an arterioscintography reading. As it is an expensive procedure, the peak cardiac enzyme level in the blood stream, which is more easily obtainable, is used as a surrogate measure of heart vascular damage (see Wittes et al., 1989). Sometimes the observed value of the response variable in the middle of an ongoing experiment is considered as the surrogate endpoint. 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 after six months is considered as a surrogate corresponding to the final outcome.

Two basic problems are studied in the literature of surrogate responses, namely (a) validation of a surrogate and (b) measurement of gain in inference using the surrogate responses. Prentice (1989) gave validation criteria for a surrogate, which is subsequently discussed by Freedman et al. (1992), Reilly and Pepe (1995), Day and Duffy (1996), Buyse and Molenberghs (1998), Buyse et al. (2000), Molenberghs et al. (2001), and Chen et al. (2007). The use of surrogate endpoints is likely to be beneficial, not only in terms of cost or time, but it gives more accuracy in the estimation of target parametric functions such as treatment difference and odds ratio. For that purpose we first look at the data structure under

* Corresponding author. Tel.: +91 33 25752818; fax: +91 33 25773104.
E-mail addresses: buddh.banerjee@iiserkol.ac.in (B. Banerjee), atanu@isical.ac.in, appubabale@gmail.com (A. Biswas).

http://dx.doi.org/10.1016/j.spl.2014.09.015
0167-7152/

consideration. In this article we are interested to measure the increment of the efficiency with the increase of the proportion of surrogate endpoints where the surrogate is presumed to be validated.

Suppose we consider the accumulated data when all of the surrogate responses are known, but only $Q\%$ of the true responses are available. Then, if we do not consider the surrogate data, we need to make inference based on $Q\%$ true responses only. In the present paper our objective is to use surrogate data efficiently (which consist of $Q\%$ bivariate data and $(100-Q)\%$ only univariate surrogate data) to improve the inference. To use $(100-Q)\%$ surrogate data efficiently one need to identify the dependency structure of true and surrogate responses based on the $Q\%$ bivariate true-and-surrogate data. A real data example which is appropriate to this situation is described in the next section.

Pepe (1992) obtained the distribution of the estimator of regression parameter when the validation sample fraction has a fixed limiting value, ρ $(=Q/100)$, say. Banerjee and Biswas (2011) established that the variance of the estimator of treatment difference is bounded for such a fixed ρ. Lin et al. (1997) measured the extent to which a biological marker is a surrogate endpoint for a clinical event and Wang and Taylor (2002) propose alternative measures of the *proportion explained* by the surrogate endpoint. Chen (2000) and Begg and Leung (2000) discussed the inferential improvement by the use of surrogate endpoints. Chen et al. (2003) introduced the concept of information recovery from surrogate endpoints 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 observations, 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 binary surrogate responses. In Section 3 we establish that the (inverse of) relative efficiency to estimate the treatment success probability by using surrogate endpoints 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, log risk ratio 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 we demonstrate our results with data example and conclude.

2. Experimental details and data structure

We consider a set up of two treatments having binary true endpoints with binary surrogates as well. Begg and Leung (2000) pointed out that for the binary endpoints the probability of concordance is an indicator of association between true and surrogate endpoints. Suppose n_A and n_B patients are allotted to the treatments A and B, respectively; but we get only m_A and m_B true endpoints along with all surrogate endpoints within the stipulated time frame or cost limit, where $m_t \ll n_t,\ t = A, B$. Denote the true and surrogate endpoints for the treatment t by Y_t and W_t, where $t = A, B$. All these endpoints 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 endpoints for treatment t. Furthermore, let us denote

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

which are the *sensitivity* and *specificity* of the 2×2 table for treatment t where the true and surrogate responses are in the two margins. Clearly it is a saturated model with full parameter space. Consequently, the success probabilities by the surrogate responses for the two treatments are

$$\begin{aligned} r_t &= P(W_t = 1) \\ &= P(W_t = 1|Y_t = 1)P(Y_t = 1) + P(W_t = 1|Y_t = 0)P(Y_t = 0) \\ &= \pi_{t1}p_t + (1-\pi_{t0})(1-p_t) \\ &= p_t(\pi_{t1} + \pi_{t0} - 1) + (1-\pi_{t0}). \end{aligned}$$

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
Only surrogate	W_{tS}	$n_t - m_t - W_{tS}$	$n_t - m_t$

where $Y_{tT} = \sum_{i=1}^{m_t} Y_{ti}$ and $W_{tT} = \sum_{i=1}^{m_t} W_{ti}$; also we denote $W_{tS} = \sum_{i=m_t+1}^{n_t} W_{ti}$ for $t = A$ and B. The notation (Y_{ti}, W_{ti}) is specifically used for denoting the response variables corresponding to the ith individual under tth treatment. If any marginal is found to be zero, it is customary to add 0.5 to each of the marginals. As an example/illustration we consider the data set analyzed by Buyse and Molenberghs (1998). This data set is obtained from a randomized clinical trial comparing an experimental treatment interferon-α, with highest dose, 6-million units daily to a corresponding placebo in the treatment of patients with age-related macular degeneration (ARMD). Patients with ARMD progressively lose vision. In the trial, a patient's visual acuity is assessed at different time points through the ability to read lines of letters on standardized vision charts. It is examined whether the loss of at least two lines of vision at 6 months (denoted as 1, and 0 otherwise) can be used