Published in final edited form as:

J R Stat Soc Ser A Stat Soc. 2013 February 1; 176(2): 603–608. doi:10.1111/j.1467-985X.2012.01052.x.

The risky reliance on small surrogate endpoint studies when planning a large prevention trial

Stuart G. Baker and Barnett S. Kramer
National Cancer Institute, Bethesda, USA

Summary

The definitive evaluation of treatment to prevent a chronic disease with low incidence in middle age, such as cancer or cardiovascular disease, requires a trial with a large sample size of perhaps 20,000 or more. To help decide whether to implement a large true endpoint trial, investigators first typically estimate the effect of treatment on a surrogate endpoint in a trial with a greatly reduced sample size of perhaps 200 subjects. If investigators reject the null hypothesis of no treatment effect in the surrogate endpoint trial they implicitly assume they would likely correctly reject the null hypothesis of no treatment effect for the true endpoint. Surrogate endpoint trials are generally designed with adequate power to detect an effect of treatment on surrogate endpoint. However, we show that a small surrogate endpoint trial is more likely than a large surrogate endpoint trial to give a misleading conclusion about the beneficial effect of treatment on true endpoint, which can lead to a faulty (and costly) decision about implementing a large true endpoint prevention trial. If a small surrogate endpoint trial rejects the null hypothesis of no treatment effect, an intermediate-sized surrogate endpoint trial could be a useful next step in the decision-making process for launching a large true endpoint prevention trial.

Keywords

Cancer prevention; Cardiovascular disease; Prentice criterion; Principal stratification; Sample size calculation; Surrogate endpoint

1. Introduction

Searching for an effective treatment to prevent a chronic disease with low incidence in a middle aged cohort, such as cancer or cardiovascular disease, is challenging because a definitive trial with a disease incidence endpoint requires a very large sample size of perhaps 20,000 or more. Before implementing such a large trial, investigators often seek evidence of treatment benefit from a small trial of perhaps 200 using a surrogate endpoint observed before the true endpoint of disease incidence. For example, a recent trial of a drug to reduce the surrogate endpoint of adenoma occurrence involved a sample size of 267 (Thompson, *et al.*, 2010). Had this study used the true endpoint of colorectal cancer incidence or mortality, the sample size could have been as large as 70,000 (Atkin, 2010). As another example, a recent trial of a drug to reduce the surrogate endpoint of the occurrence of bronchial dysplasia involved a sample size of 100 (Lam, *et al.*, 2004). Had this study used the true endpoint of lung cancer incidence, the sample size could have been as large as 30,000 (The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994). These surrogate endpoint trials have adequate power to detect an effect of treatment on a surrogate endpoint,

and their small sample sizes are touted as an important advantage (Psaty *et al.*, 1999). But is this the proverbial case of a free lunch that is not really free?

To answer this question, it is helpful to consider the primary goal of using a surrogate endpoint and the data available to achieve this goal. The following discussion of the use of surrogate endpoints is not a summary of an extensive statistical literature (e.g., Walle and Weir, 2006; Lassere, 2008) but focuses on key points in study design and implications.

With treatment trials, usually the primary goal of using a surrogate endpoint is extrapolation, namely drawing conclusions about the effect of treatment on a true endpoint while shortening the duration of the trial. In this setting data are typically available from at least one historical trial with the same surrogate and true endpoints associated with the trial of interest. Typically a model constructed from these historical data is used to predict the effect of treatment on the true endpoint based on the surrogate endpoint in the trial of interest (e.g. Baker *et al.* 2012).

With the prevention trials discussed here, usually the primary goal of using a surrogate endpoint is to draw conclusions about the effect of treatment on a true endpoint using a much smaller sample size than with the true endpoint. (A secondary goal is shortening the duration of the study). In terms of available data, typically there are no historical trials with the same surrogate and true endpoints as associated with the trial of interest. With no data for predicting treatment effect on true endpoint, the focus is on hypothesis testing in order to justify a much larger definitive trial. In the hypothesis testing framework, Prentice (1989) defined a valid surrogate endpoint as a surrogate endpoint satisfying what we call the *Extrapolation Assumption*, namely rejecting the null hypothesis of no treatment effect on a surrogate endpoint in favor of a beneficial treatment effect on the surrogate endpoint implies rejecting the null hypothesis of no treatment effect on a true endpoint in favor of a beneficial treatment effect on the true endpoint.

Using three models relating surrogate and true endpoints, we show that the link between the Extrapolation Assumption and the size of the surrogate endpoint trial explains why a small surrogate endpoint trial is particularly unreliable for drawing conclusions about the effect of treatment on the true endpoint.

2. Binary surrogate endpoint: mixture model

Let $S = 0, 1$ and $T = 0, 1$ denote binary surrogate and true endpoints, respectively, where outcome 0 (1) is unfavorable (favorable). For example $T = 0$ (1) is incidence (no incidence) of lung cancer and $S = 0$ (1) is occurrence (non-occurrence) of bronchial dysplasia. Let $Z = 0$ (control), 1 (experimental) denote the randomization group. Let $p_z = \text{pr}(S=1 | Z=z)$, $f_z = \text{pr}(T=1 | Z=z)$, $b_s = \text{pr}(T=1 | S=s, Z=0)$, and $c_s = \text{pr}(T=1 | S=s, Z=1) - \text{pr}(T=1 | S=s, Z=0)$. The probabilities of true endpoint are mixtures

$$f_0 = p_0 b_1 + (1 - p_0) b_0 \quad (1)$$

$$f_1 = p_1 (b_1 + c_1) + (1 - p_1) (b_0 + c_0). \quad (2)$$

Equations (1) and (2), which define the mixture model in Baker *et al.* (2012), imply $f_1 - f_0 = (p_1 - p_0) (b_1 - b_0) + d_{MIX}$, where $d_{MIX} = p_1 c_1 + (1 - p_1) c_0$.

The goal is to reject $f_1 = f_0$ in favor of $f_1 > f_0$. For a reasonable surrogate endpoint, $b_1 - b_0 > 0$. Therefore

$$p_1 - p_0 > 0 \text{ implies } f_1 - f_0 > d_{MIX}. \quad (3)$$

The Extrapolation Assumption says that $p_1 - p_0 > 0$ (namely rejecting the null hypothesis of no treatment effect on a surrogate endpoint) implies $f_1 - f_0 > 0$ (namely rejecting the null hypothesis of no treatment effect on a true endpoint); this is a special case of Equation (3) in which $d_{MIX} = 0$. Thus if an investigator believes the Extrapolation Assumption holds, but in reality $d_{MIX} < 0$, it is possible to conclude there is beneficial effect of treatment on the true endpoint when, in fact, there is a detrimental effect of treatment on the true endpoint equal to d_{MIX} . To put the magnitude of d_{MIX} in perspective relative to the effect of treatment, we compute the *relative error* (as a percent), namely $RE_{mix} = 100 d_{MIX} / (f_1 - f_0)$. The requirement for the Extrapolation Condition that $d_{MIX} = 0$ implies $c_0 = c_1 = 0$, which is called the *Prentice Criterion*, namely the probability of true endpoint given surrogate endpoint does not depend on randomization group. See also Buyse and Molenberghs (1998).

For a true endpoint trial, a two-sided type I error of 5%, and power of 90%, a standard formula for the sample size of each equal-sized randomization group (Halperin, et al. 1968) is

$$\text{Size}(f_0, f_1) = \left[1.96 \{f_1(1 - f_1) + f_0(1 - f_0)\}^{1/2} + 1.28 \{2 f_{avg}(1 - f_{avg})\}^{1/2} \right]^2 / (f_1 - f_0)^2. \quad (4)$$

where $f_{avg} = (f_0 + f_1)/2$. The sample size for the surrogate endpoint trial with a two-sided type I error of 5% and power of 90% is $\text{Size}(\rho_0, \rho_1)$ computed under the Prentice Criterion.

Consider a realistic example with $f_0 = 0.003$ and $f_1 = 0.004$ (Table 1). Realistic values of parameters under the Prentice Criterion can be obtained using proportional probabilities of endpoints, namely $p_z = R f_z$, which implies $b_0 = 0$ and $b_1 = 1/R$. For $R = 100$, a sample size of 73300 for a true endpoint trial is reduced to 480, but the relative error for a very small deviation from the Prentice Criterion of $c_1 = -0.002$ is -80% , a possibility of great concern. If $1 - p_z$ is close to 1, $\text{Size}(R \rho_0, R \rho_1) / \text{Size}(\rho_0, \rho_1) \approx 1/R$. For a deviation from the Prentice Criterion in only c_1 (with $c_0 = 0$), the relative error is $RE_{MIX}(\rho_1) = 100 p_1 c_1 / (f_1 - f_0)$, which implies $RE_{MIX}(R \rho_1) / RE_{MIX}(\rho_1) = R$. Thus, regardless of the values of f_0 and f_1 , if the sample size decreases by a factor of approximately R , the relative error increases by a factor of R , in agreement with Table 1.

3. Binary surrogate endpoint: principal stratification model

The principal stratification model for surrogate endpoints (Frangakis and Rubin, 2002) was originally formulated to estimate causal effects. Here we discuss its implications for hypothesis testing. Let S^* denote the four principal strata: A (always), when $S = I$ regardless of randomization assignment, C (consistent) when $S = z$, I (inconsistent) when $S = I - z$, N (never) when $S = 0$ regardless of randomization assignment. Let $p_{s^*} = \text{pr}(S^* = s^*)$, $b_{s^*} = \text{pr}(T=1 | S^* = s^*, Z=0)$, $h_{s^*} = \text{pr}(T=1 | S^* = s^*, Z=1) - \text{pr}(T=1 | S^* = s^*, Z=0)$. By definition, $p_1 = p_A + p_C$ because $S=1$ for $Z=1$ only in principal strata A and C . Similarly $p_0 = p_A + p_I$ because $S=1$ for $Z=0$ only in principal strata A and I . Consequently $p_1 - p_0 = p_C - p_I$. Combining this result with

$$f_0 = p_A b_A + p_C b_C + p_I b_I + p_N b_N, \quad (5)$$

$$f_1 = p_A (b_A + h_A) + p_I (b_I + h_I) + p_C (b_C + h_C) + p_N (b_N + h_N), \quad (6)$$

gives $f_1 - f_0 = (p_1 - p_0) h_C + d_{PS}$, where $d_{PS} = h_A p_A + h_N p_N + (h_I - h_C) p_I$. The relative error is $RE_{PS} = 100 d_{PS} / (f_1 - f_0)$. Here the Extrapolation Assumption, $p_1 - p_0 > 0$ implies $f_1 - f_0 > 0$, requires $d_{PS} = 0$. An appealing scenario for $d_{PS} = 0$ is what we call the *PS Criterion*: $h_A = h_N = 0$ and $p_I = 0$, namely the probability of true endpoint for principal strata A and N depends only on the level of the surrogate endpoint and not randomization group and there is no person for whom the level of surrogate endpoint is “inconsistent” with the randomization group. The PS Criterion is analogous to the identifiability requirement in some principal stratification models (e.g. Baker *et al.*, 2011).

The sample size formula for a true endpoint trial is $Size(f_0, f_1)$. The sample size for the surrogate endpoint trial with a two-sided type I error of 5% and power of 90% is $Size(p_0, p_1)$ computed under the PS Criterion.

Again let $f_0 = 0.003$ and $f_1 = 0.004$. Realistic values of parameters under the PS Criterion can be obtained using $p_z = R f_z$, which implies $p_A = f_0 R$, $p_C = (f_1 - f_0) R$, and $p_N = 1 - f_1 R$. For $R = 100$, a sample size of 73300 for a true endpoint trial is reduced to 480 (Table 1), and the relative error for a very small deviation from the PS Criterion of $h_A = -0.002$ is -60%, a possibility of great concern. If $1 - p_z$ is close to 1, $Size(R p_0, R p_1) / Size(p_0, p_1) \approx 1/R$. For the deviation from the PS Criterion in only h_A (with $h_N = p_I = 0$) the relative error is $RE_{PS}(p_0) = 100 h_A p_A / (f_1 - f_0) = h_A p_0 / (f_1 - f_0)$, so $RE_{PS}(R p_0) / RE_{PS}(p_0) = R$. Thus, as with the mixture model, if the sample size decreases by a factor of approximately R , the relative error increases by a factor of R , in agreement with Table 1.

4. Continuous surrogate endpoint: model for the mean

To investigate small surrogate endpoint trials with a continuous surrogate endpoint, let s_z denote the mean value of a continuous surrogate endpoint for randomization group z , and let $t_z = \logit(f_z)$. Also let $(\sigma_T)^2$ and $(\sigma_S)^2$ denote the variance for a sample size of 1 of $(t_1 - t_0)$ and $(s_1 - s_0)$, respectively. Based on the delta method, $(\sigma_T)^2 = 1 / \{f_0 (1 - f_0)\} + 1 / \{f_1 (1 - f_1)\}$. A simple linear model with σ_T and σ_S as scale factors,

$$(t_0 / \sigma_T) = a_0 + b (s_0 / \sigma_S) \quad (7)$$

$$(t_1 / \sigma_T) = a_1 + (b + c) (s_1 / \sigma_S) \quad (8)$$

implies $(t_1 - t_0) / \sigma_T = b (s_1 - s_0) / \sigma_S + d_{MEAN}$, where $d_{MEAN} = c (s_1 / \sigma_S) + a_1 - a_0$. The relative error is $RE_{MEAN} = 100 d_{MEAN} / \{(t_1 - t_0) / \sigma_T\}$. Under this model $s_1 - s_0 > 0$ implies $t_1 - t_0 > d_{MEAN}$. Here the Extrapolation Assumption, $s_1 - s_0 > 0$ implies $t_1 - t_0 > 0$, requires $d_{MEAN} = 0$. An appealing scenario for $d_{MEAN} = 0$ is what we call the *Prentice Criterion for the Mean*: $a_0 = a_1$ and $c = 0$, which implies the same effect of the mean surrogate endpoint on the true endpoint for each randomization group.

For the true endpoint trial with a two-sided type I error of 5%, and power of 90%, the sample size of each equal-sized randomization group is $Size^*(t_0, t_1) = (1.96 + 1.28)^2 (\sigma_T)^2 / (t_1 - t_0)^2$. The sample size for the surrogate endpoint trial with a two-sided type I error of 5% and power of 90% is $Size^*(s_0, s_1)$ computed under the Prentice Criterion for the Mean, which we can write as $Size^*(b) = Size^*(t_0, t_1) / b^2$.

Again consider $f_0 = 0.003$ and $f_1 = 0.004$. The sample size for the true endpoint trial is 73,800. To obtain realistic parameter values, we consider sample sizes computed under the other models. For the sample size of 480 (Table 1), the relative error arising from a small deviation from the Prentice Criterion for the Mean of $c = 0.002$ is -46.1%, yet another possibility of great concern. Note that $Size^*(b/k) / Size^*(b) = k^2$. For the deviation from

Prentice Criterion for the Mean involving only c (with $a_0 = a_1$), $RE_{MEAN}(b) = 100 t_1 \{ c / (b + c) \} / (t_1 - t_0)$, so $RE_{MEAN}(b/k) / RE_{MEAN}(b) = (b + c)/(b/k + c) \approx k$ for small c . Thus, if sample size decreases by a factor of k^2 , the relative error increases by a factor of approximately k , in agreement with Table 1.

5. Discussion

The search for treatments to prevent cancer or cardiovascular disease involves preliminary evaluations using small surrogate endpoint trials, and as a practical matter this trend will likely continue in the genomic era in order to handle the likely explosion in potential hypotheses to be tested. Although a small surrogate endpoint trial is typically conducted, it has a greater potential than a large surrogate endpoint trial for a misleading conclusion that treatment has a beneficial effect on true endpoint. This misleading conclusion could lead to a faulty decision about implementing a definitive trial with a true endpoint. The implications regarding expenditure of resources are enormous.

The focus here is on an incorrect conclusion to implement a large prevention trial after rejecting the null hypothesis in the surrogate endpoint study. It is also possible to draw an incorrect conclusion of not implementing a large prevention trial after not rejecting the null hypothesis in the surrogate endpoint study. However this latter incorrect conclusion is of limited interest because the beneficial treatment effect would not likely be large and researchers are looking for a large beneficial treatment effect to make a large prevention trial worthwhile.

If a small surrogate endpoint trial indicates a promising treatment, investigators should next investigate an intermediate-sized surrogate endpoint trial (with a more frequently occurring surrogate endpoint) before jumping to a very large, resource-intensive, prevention trial with a true endpoint of a rare disease. Of course the intermediate-sized surrogate endpoint trial is no guarantee of drawing a correct conclusion. Therefore other sources of evidence, such as observational studies and animal testing, need to be considered before implementing a large true endpoint prevention trial. Also surrogate endpoint trials typically do not provide information about multiple endpoints and long-term side effects, which is another reason for caution.

Acknowledgments

This research was supported by the National Institutes of Health.

References

- Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JM, Parkin DM, Wardle J, Duffy SW, Cuzick J, UK Flexible Sigmoidoscopy Trial Investigators. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*. 2010; 8:1624–1633. [PubMed: 20430429]
- Baker SG, Sargent DJ, Buyse M, Burzykowski T. Predicting treatment effect from surrogate endpoints and historical trials: an extrapolation involving probabilities of a binary outcome or survival to a specific time. *Biometrics*. 2012; 68:248–257. [PubMed: 21838732]
- Baker SG, Lindeman KS, Kramer BS. Clarifying the role of principal stratification in the paired availability design. *International Journal of Biostatistics*. 2011; 7:1.
- Baker SG, Lindeman KS. The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*. 1994; 13:2269–2278. [PubMed: 7846425]
- Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998; 54:1014–1029. [PubMed: 9840970]

- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
- Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *Journal of Chronic Diseases*. 21:13–24. [PubMed: 5646357]
- Lam, S, leRiche J. C.; McWilliams, A.; Macaulay, C.; Dyachkova, Y.; Szabo, E.; Mayo, J.; Schellenberg, R.; Coldman, A.; Hawk, E.; Gazdar, A. A randomized phase IIb trial of pulmicort turbuhaler (budesonide) in people with dysplasia of the bronchial epithelium. *Clinical Cancer Research*. 2004; 10:6502–11. [PubMed: 15475437]
- Lassere MN. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Statistical Methods in Medical Research*. 2008; 17:303–340. [PubMed: 17925313]
- Psaty BM, Weiss NS, Furberg CD, Koepsell TD, Siscovick DS, Rosendaal FR, Smith NL, Heckbert SR, Kaplan RC, Lin D, Fleming TR, Wagner EH. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. *Journal of the American Medical Association*. 1999; 282:786–790. [PubMed: 10463718]
- Prentice RL. Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine*. 1989; 8:431–430. [PubMed: 2727467]
- The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The Effect of vitamin E and beta Carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine*. 1994; 330:1029–35. [PubMed: 8127329]
- Thompson PA, Wertheim BC, Zell JA, Chen WP, McLaren CE, LaFleur BJ, Meyskens FL, Gerner EW. Levels of rectal mucosal polyamines and prostaglandin E2 predict, ability of DFMO and Sulindac to prevent colorectal adenoma. *Gastroenterology*. 2010; 139:797–805. [PubMed: 20538001]
- Weir CJ, Walle RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine*. 2006; 25:183–203. 431–430. [PubMed: 16252272]

Table 1

Relative errors and sample sizes.

Formulation	Key parameter	Sample size per randomization group		Deviation from Extrapolation Assumption	Relative Error (%)
		True endpoint	Surrogate endpoint		
Binary: mixture	$R = 0.01$	73300	480	$c_1 = -0.002$ instead of 0	-80.0
	$R = 0.10$		7100		-8.0
	$R = 1.00$		73300		-0.8
Binary: principal stratification	$R = 0.01$	73300	480	$h_A = -0.002$ instead of 0	-60.0
	$R = 0.10$		7100		-6.0
	$R = 1.00$		73300		-0.6
Continuous: Model for Mean	$b = 0.08$	73800	480	$c = 0.002$ instead of 0	-46.1
	$b = 0.31$		7100		-12.2
	$b = 1.00$		73300		-3.8

The relative error is the percentage change in the effect of treatment on the true endpoint (under the deviation from the Extrapolation Assumption) that is consistent with rejecting the null hypothesis of no effect of treatment on the surrogate endpoint (when incorrectly making the Extrapolation Assumption). Computations are based on probabilities of true endpoint of $t_0 = 0.003$ and $t_1 = 0.004$.