BMC Medical Research Methodology

BioMed Central

Research article

Open Access

A perfect correlate does not a surrogate make

Stuart G Baker*[1] and Barnett S Kramer[2]

Address: [1]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, USA and [2]Office of Disease Prevention, National Institutes of Health, USA

Email: Stuart G Baker* - sb16i@nih.gov; Barnett S Kramer - KramerB@OD.NIH.GOV

* Corresponding author

Published: 09 September 2003

BMC Medical Research Methodology 2003, **3**:16

Received: 10 July 2003

Accepted: 09 September 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/16

Abstract

Background: There is common belief among some medical researchers that if a potential surrogate endpoint is highly correlated with a true endpoint, then a positive (or negative) difference in potential surrogate endpoints between randomization groups would imply a positive (or negative) difference in unobserved true endpoints between randomization groups. We investigate this belief when the potential surrogate and unobserved true endpoints are perfectly correlated within each randomization group.

Methods: We use a graphical approach. The vertical axis is the unobserved true endpoint and the horizontal axis is the potential surrogate endpoint. Perfect correlation within each randomization group implies that, for each randomization group, potential surrogate and true endpoints are related by a straight line. In this scenario the investigator does not know the slopes or intercepts. We consider a plausible example where the slope of the line is higher for the experimental group than for the control group.

Results: In our example with unknown lines, a *decrease* in mean potential surrogate endpoints from control to experimental groups corresponds to an *increase* in mean true endpoint from control to experimental groups. Thus the potential surrogate endpoints give the wrong inference. Similar results hold for binary potential surrogate and true outcomes (although the notion of correlation does not apply). The potential surrogate endpointwould give the correct inference if either *(i)* the unknown lines for the two group coincided, which means that the distribution of true endpoint conditional on potential surrogate endpoint does not depend on treatment group, which is called the Prentice Criterion or *(ii)* if one could accurately predict the lines based on data from prior studies.

Conclusion: Perfect correlation between potential surrogate and unobserved true outcomes within randomized groups does not guarantee correct inference based on a potential surrogate endpoint. Even in early phase trials, investigators should not base conclusions on potential surrogate endpoints in which the only validation is high correlation with the true endpoint within a group.

Background

A potential surrogate endpoint is an endpoint obtained sooner, at less cost, or less invasively than the true endpoint of interest. When using a potential surrogate endpoint, one would like to make the same inference as if one had observed a true endpoint (i.e. a health outcome).

Fleming and DeMets [1] and the Biomarker Definitions Working Group [2] gave various examples where preliminary inference based on a potential surrogate endpoint was contradicted by later studies using important health outcomes. As another example, it has long been assumed that postmenopausal hormone replacement therapy (HRT) with estrogen and progestin would decrease the risk of cardiac disease in women-in part due to the facts that (1) hormonal therapy lowers serum cholesterol, and (2) people with low cholesterol generally have a lowered risk of cardiac disease. However direct assessment of this hypothesis in a randomized placebo controlled trial showed that HRT actually *increased* the incidence of cardiac events [3].

Fleming and DeMets [1] wrote "A correlate does not a surrogate make" and said that "it is a common misconception that if an outcome is correlated (that is correlated with true clinical outcome) it can be used as a valid surrogate endpoint..". They added that arequirement for a valid surrogate is that" the effect of the intervention on the potential surrogate endpoint predicts the effect on clinical outcome---a much stronger condition than correlation." Using schema for causal pathways, Fleming and DeMets [1] showed why a potential surrogate endpoint can fail to provide correct inference about the true endpoint. However Fleming and DeMets [1] did not show why perfect correlation of potential surrogate and true endpoints is insufficient for correct prediction of the true endpoint. A reader of Fleming and DeMets [1] might incorrectly conclude that the failure of potential surrogate endpoints only occurs with small or moderate correlations between potential surrogate and true endpoints, but not with perfect correlation within randomized groups. More specifically, some investigators believe that that if they have evidence of a very high correlation between the potential surrogate and the true outcome in each treatment group in a previous study, they can make reliable inference about true outcome in a new study with only a surrogate endpoint. The main purpose of this paper is to show graphically that even a perfect correlate within a randomized group does not a valid surrogate make.

Methods

To investigate the validity of a potential surrogate endpoint that is perfectly correlated with true outcomes within randomized groups, we created the graphic in Figure 1 for measured outcomes. The graphic applies to a hypothetical randomized trial, where E stands for experimental group and C stands for control group. The horizontal axis is the potential surrogate endpoint and the vertical axis is the unobserved true endpoint. Assuming perfect correlation, the individual data points for potential surrogate and unobserved true endpoints lie on straight lines for groups E and C.

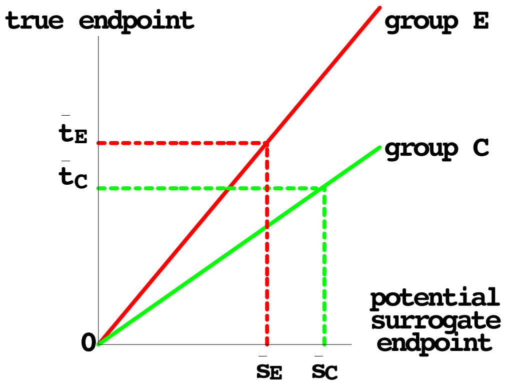

Figure 1
Graphical depiction of incorrect inference based on surrogage endpoints. The graph shows perfectly correlated results (namely a straight line) for the relationship between surrogate and true outcomes for a control group C and experimental group E. The mean surrogate outcome in the E group $\bar{s}_E$ is smaller than the mean surrogate outcome in the C group $\bar{s}_C$. However the mean true outcome in the E group $\bar{t}_E$ is larger than the mean true outcome in the C group, $\bar{t}_C$, yielding the opposite conclusion for the effect of experimental intervention.

In this simple example the unobserved true endpoint is proportional to the potential surrogate endpoint, so the intercept for both lines is zero. However qualitatively similar results would hold when the two lines have different intercepts. In our example the onlydifference between the lines for groups E and C is that the slopes differ. In particular, the slope of the line relating potential surrogate and true endpoints for group E is higher than that for group C. To graphically find the mean value of a true endpoint corresponding to the mean value of potential surrogate endpoint, one draws a vertical line from the mean value of the potential surrogate endpoint to the line relating potential surrogate and true endpoint, and then a horizontal line leftward to the axis for the true endpoint.

The graphic also applies when the potential surrogate and true outcomes are binary. Of course with binary endpoints the notion of perfect correlation does not apply. However, with binary endpoints, one obtains straight lines, so the graphic is applicable. For example, suppose the potential surrogate endpoint is the presence or absence of adenoma and the true endpoint is the presence or absence of colorectal cancer. In that case the horizontal axis is the fraction of subjects with adenomas and the vertical axis is the unobserved fraction of subjects who would get colorectal cancer. For each randomization group there is a line relating the fraction of subjects with adenoma to the unobserved fraction with colorectal cancer. (Each line is constructed by connecting the point representing the proportion with the true endpoint when the surrogate endpoint is 0 with the point representing the proportion with true endpoint when the surrogate endpoint is 1). The lines in Figure 1 represent one example with a binary surrogate endpoint.

Suppose that an investigator only knows from prior studies that the potential surrogate and true endpoints are perfectly correlated within randomization group (and does not know the slopes or intercepts). Or suppose the surrogate and true endpoints are binary, so there are two straight lines (but with unknown slopes or intercepts), one for each randomization group. Will the use of a potential surrogate endpoint to replace the unobserved true endpoint give qualitatively the correct results? In other words, will a decrease (increase) in the mean potential surrogate endpoint or the fraction with the surrogate endpoint necessarily imply a decrease (increase) in the mean true endpoint or the fraction with the true endpoint?

Results

In the graphic in Figure 1, the mean value for the potential surrogate endpoint for group E, denoted by $\bar{s}_E$, is smaller than the mean value of the potential surrogate endpoint for group C, denote by $\bar{s}_C$. Because of the perfect correlation between potential surrogate and true endpoints, one might naively think that the mean for the true endpoint for group E, denoted by $\bar{t}_E$, would be *smaller* than the mean value of the true endpoint for group C, denote by $\bar{t}_C$. However if we examine the graphic the opposite is true. Using the graphic to go from potential surrogate value to the true value (with a vertical line upwards and a horizontal line to the left), we see that, in fact, $\bar{t}_E$, is *larger* than $\bar{t}_C$, so the naive conclusion is erroneous.

With binary data, $\bar{s}_E$ and $\bar{s}_C$ represent the fraction of subjects in the experimental and control groups with the surrogate endpoint, and $\bar{t}_E$ and $\bar{t}_C$ represent the fraction of subjects in the experimental and control groups with the true endpoint. Based on this graphic it is possible that a *decrease* in the fraction of subjects with adenoma would correspond to an *increase* in the fraction of subjects with colorectal cancer.

One could create a similar graphic that shows that no change in the surrogate endpoint corresponds to either a decrease or increase in the true endpoint, or that an increase in the surrogate endpoint could lead to a decrease in the true endpoint.

Discussion

Plausibility of Figure 1

We showed graphically that perfect correlation does not guarantee correct inference when a potential surrogate endpoint replaces a true endpoint. The underlying reason is that the line predicting true endpoint from potential surrogate endpoint has a sufficiently different slope for each randomization group to make a substantial difference in the conclusion. In one possible scenario the intervention reduces the value of the surrogateendpoint that is observed without affecting the true endpoint, thereby increasing the slope.

With binary outcomes, different slopes can readily arise because of unobserved heterogeneity in the potential surrogate endpoint. As an example consider adenoma (yes or no) as a potential surrogate endpoint for the true outcome of colorectal cancer (yes or no). Two recent randomized trials [4,5] showed that aspirin versus placebo lowers the risk of adenomas. Can one conclude that aspirin lowers the risk of colorectal cancer? An editorial on these trials [6] states "given the belief that the development of most colorectal cancers follows a sequence leading from adenoma to carcinoma, a clinical trial in which aspirin reduced the rate of recurrence of adenomas might make a compelling case for its effectiveness." However we disagree (and the editorial later comes to a similar conclusion). Under the single pathway hypothesis, if the probability of adenoma is zero, the probability or colorectal cancer is zero regardless of the intervention (as there is no other way to get colorectal cancer). Thus, in terms of our graphic, the single pathway hypothesis implies that the intercepts of the lines for each group are 0, as in Figure 1. However the slopes can differ substantially due to heterogeneity of adenomas, for example in a spectrum of histological types and sizes [7].

To better understand the role of heterogeneity, we follow Schatzkin and Gail [8], and suppose that there are two types of adenomas: "bad" adenomas that have the potential to develop into colorectal cancer and "innocent" adenomas that do not. Let π_z denote the probability an adenoma in randomization group z is "bad." Let ϕ_z denote the probability of colorectal cancer arising from a "bad" adenoma in randomization group z. Also let ω_z denote the probability of any adenoma in group z. The larger slope in the experimental than the control group in Figure 1 would occur if $\phi_E\pi_E > \phi_C\pi_C$. The leftward shift of the vertical line in Figure 1 would occur if $\omega_E < \omega_C$.

Such a situation is quite plausible, as illustrated in a randomized trial of finasteride versus placebo [9] where the potential surrogate endpoint was the probability of prostatecancer. A true definitive endpoint would be the probability of death from prostate cancer. In this trial, heterogeneity was observed in the form of high-grade prostate cancer versus other histological grades of prostate cancer. Relative to the placebo group, the finasteride group consisted of a greater fraction of men with high-grade prostate cancer ($\pi_E > \pi_C$) but a smaller fraction with any prostate cancer ($\omega_E < \omega_C$). Because individuals with high-grade prostate cancer generally have a greater risk of prostate cancer mortality, we have $\phi_E > \phi_C$. If the risk of prostate cancer mortality with other histological grades of prostate cancer is minimal, the situation is mathematically similar to the aforementioned hypothetical example with "bad" and "innocent" adenomas, except that that the fraction "bad" is observed. There is a greater slope with the finasteride group ($\phi_E\pi_E > \phi_C\pi_C$) and a smaller value fraction with the surrogate with the finasteride group ($\omega_E < \omega_C$), which corresponds to Figure 1.

Graphical Representation of the Prentice Criterion

For valid hypothesis testing based on a surrogate endpoint that *replaces* a true endpoint, Prentice developed three criteria [10]. The major one, subsequently called the Prentice Criterion, is that the distribution of true endpoint given the potential surrogate endpoint does not depend on treatment group [10]. Our graphic shows that if the potential surrogate endpoint is a perfect correlate for a true endpoint (even if the slopes and intercepts of the lines were unknown) *and* if the Prentice Criterion holds, one would obtain the correct inference about the true endpoint based on the potential surrogate endpoint. Graphically, the Prentice Criterion implies that the lines for groups E and C coincide, so a decrease in the mean potential surrogate endpoint would always translate into a decrease in the mean true endpoint. Wang and Taylor [11] developed a similar graphic to help explain their proposed statistic, the proportion of treatment effect summarized by the potential surrogate, which indicates the appropriateness of the Prentice Criterion.

Inference Without the Prentice Criterion

Other approaches to inference with surrogate endpoints involve *predicting* the true endpoint conditional on the surrogate endpoint (and using estimates based on data from previous studies). This use of potential surrogate endpoints to predict true endpoints differs from the use of auxiliary variables to predict true endpoint. An auxiliary variable is a variable that occurs after randomization and before a true endpoint that is missing in *some but not all* subjects. (See [12] and references therein which discuss the role of auxiliary variables in increasing efficiency or reducing bias.) In contrast a potential surrogate endpoint occurs before a true endpoint that is missing in *all* subjects.

If one could accurately predict the slopes and intercepts of lines in Figure 1 based on data from previous studies, one could obtain the correct inference even if the Prentice Criterion did not hold (i.e. even if the lines did not coincide). For example, in Figure 1, if the slopes and intercepts of the lines were accurately predicted, one could correctly predict that the experimental intervention increases the mean value of the true endpoint despite the decrease in the mean value of the potential surrogate endpoint (and in fact obtain estimates and confidence intervals for the predicted increase in the true endpoint). Unfortunately, this situation is infrequent. In practice, sufficiently accurate prediction of the lines based on previous data is difficult because of sampling variability in the estimates of the intercepts and slopes of each line and because each previous study will likely generate a different line, even without sampling variability, due to differences in interventions. (Although in practice, the only relevant part of the lines occurs at the mean values of the surrogate endpoint in the new study).

Another approach for predicting true endpoint from a potential surrogate endpoint is the "meta-analytic" approach [13,14]. The meta-analytic approach is not reflected in Figure 1 because it does not involve the distribution of the true outcome conditional on the potential surrogate endpoint (i.e. the slanted lines). Instead each previous trial generatestwo regressions: one for the effect of intervention on potential surrogate endpoint and one for the effect of intervention on the true endpoint. The coefficients for these two regressions are treated as random variables with a joint distribution over all trials. The estimated parameters from this joint distribution are used to predict the difference in mean true endpoints in a new trial given the mean values of the potential surrogate endpoints in the new trial. Unfortunately, there are infrequently sufficient data to use this method routinely, and confidence intervals can be very wide due to between-study variation [13].

A third approach for predicting true endpoints from a potential surrogate endpoint in a randomized trial is a counterfactual approach based on the potential surrogates that would occur, if contrary to fact, individuals were randomized to a different group [15]. Estimates come from a previous study but this could be extended to multiple previous studies. Because counterfactual outcomes are not observed, additional assumptions are needed for inference.

In all these approaches there is a fundamental assumption that the relationship between the potential surrogate and true endpoints in previous studies is very similar to the relationship in the new study under investigation. Besides accounting for the variability in this relationship (in addition to sampling variability), one needs to restrict the previous studies to those involving similar interventions although, as discussed below, that is not a guarantee of valid inference.

Additional Caveats with Potential Surrogate Endpoints

The use of surrogate endpoints is particularly attractive for studies of complex chronic disease since occurrence of the true endpoint may take years. However, it is precisely because of the complexity of the diseases that assessment of potential surrogate endpoints is so difficult. There are likely to be multiple causal pathways to the true disease endpoint. Different interventions may exert their biologic effects on different pathways.

This is why it is particularly hazardous to use even an "established " surrogate endpoint (or a potential surrogate endpoint "validated" via multiple previous studies) for one class of drug to assess another class of drugs. For example, the statin class of drugs lowers serum cholesterol and lowers cardiovascular event rates, including mortality. However HRT with combined estrogen plus progestin lowers serum cholesterol but *increases* cardiovascular event rates. Presumably HRT exerts it harm via another (dominant) causal pathway, such as the induction of a hypercoagulable sate in the coronary arteries. New molecular insights into pathogenesis suggest that cancer pathogenesis is at least as complex as this situation, involving numerous causal pathways.

Another cautionary note is important. If an intervention induces harmful side effects, it is risky to draw conclusions from the potential surrogate endpoint based only inference regarding the true endpoint. There may be harms that occur after the time the potential surrogate endpoint is observed that are not well predicted by the potential surrogate endpoint. Under these circumstances, even if the two lines in Figure 1 were superimposed, acceptance of the potential surrogate endpoint could still lead to harm.

Conclusion

Sometimes potential surrogate endpoints are justified because they are highly correlated with the true endpoint in other studies. As illustrated here, even with known perfect correlation within randomized groups, one cannot rely on the potential surrogate endpoint for valid inference about the true endpoint, as even the direction of the effect could be the opposite with true and potential surrogate endpoints. Thus, even in preliminary trials, investigators should not base conclusions on potential surrogate endpoints in which the only validation is high correlation with the true endpoint.

Authors' Contribution

SGB wrote the initial draft. BSK made substantial improvements to the manuscript.

References

1. Fleming TR and DeMets DL: **Surrogate end points in clinical trials: Are we being misled?** *Annals of Internal Medicine* 1996, **125:**605-613.
2. Biomarkers Definition Working Group: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.** *Clinical Pharmacology and Therapeutics* 2001, **69:**89-95.
3. Writing Group for the Women's Health Initiative Investigators: **Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative Randomized Controlled Trial.** *Journal of the American Medical Association* 2002, **288:**2321-333.
4. Sandier RS, Halabi S, Baron JA, Budinger S, Paskett E, Keresztes R, Petrelli N, Pipas JM, Karp DD, Loprinzi CL, Steinbach G and Schilsky R: **A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer.** *New England Journal of Medicine* 2003, **348:**883-890.
5. Baron JA, Cole BF, Sandier RS, Haile RW, Ahnen D, Bresalier R, McKeown-Eyssen G, Summers RW, Rothstein R, Burke CA, Snover DC, Church TR, Allen JI, Beach M, Beck GJ, Bond JH, Byers T, Greenberg ER, Mandel JS, Marcon N, Mott LA, Pearson L, Saibil F and van Stolk RU: **A randomized trial of aspirin to prevent colorectal adenomas.** *New England Journal of Medicine* 2003, **348:**891-899.
6. Imperiale TF: **Aspirin and the prevention of colorectal cancer.** *New England Journal of Medicine* 2003, **348:**879-880.
7. Levin B: **Potential pitfalls in the use of surrogate endpoints in colorectal adenoma chemoprevention.** *Journal of the National Cancer Institute* 2003, **95:**697-698.
8. Schatzkin A and Gail M: **The promise and peril of surrogate endpoints in cancer research.** *Nature Reviews Cancer* 2002, **2:**1-9.
9. Thompson IM, Goodman PJ, Tangen CM, Lucia MS, Miller GJ, Ford LG, Lieber MM, Cespedes RD, Atkins JN, Lippman SM, Carlin SM, Ryan A, Szczepanek CM, Crowley JJ and Coltman CA: **The influence of finasteride on the development of prostate cancer.** *New England Journal of Medicine* 2003, **349:**215-224.
10. Prentice RL: **Surrogate endpoints in clinical trials: Definitions and operational criteria.** *Statistics in Medicine* 1989, **8:**431-430.
11. Wang Y and Taylor JMG: **A measure of the proportion of treatment effect explained by a surrogate marker.** *Biometrics* 2002, **58:**803-812.
12. Baker SG: **Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance.** *Journal of the American Statistical Association* 2000, **95:**43-50.
13. Gail MH, Pfeiffer R, Houwelingen HC and Carroll RJ: **On meta-analytic assessment of surrogate outcomes.** *Biostatistics* 2001, **3:**231-246.
14. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T and Alonso A: **Statistical challenges in the evaluation of surrogate endpoints in randomized trials.** *Controlled Clinical Trials* 2002, **23:**607-625.
15. Frangakis CE and Rubin DB: **Principal stratification in casual inference.** *Biometrics* 2002, **58:**21-29.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/16/prepub