Published in final edited form as:

Biometrics. 2012 March ; 68(1): 248–257. doi:10.1111/j.1541-0420.2011.01646.x.

Predicting treatment effect from surrogate endpoints and historical trials: an extrapolation involving probabilities of a binary outcome or survival to a specific time

Stuart G. Baker,

National Cancer Institute, EPN 3131, 6130 Executive Blvd MSC 7354, National Cancer Institute, Bethesda, MD 20892-7354, USA

Daniel J. Sargent,

Mayo Clinic, Rochester USA

Marc Buyse, and

IDDI, Louvain-la-Neuve, and Hasselt University, Diepenbeek, Belgium

Tomasz Burzykowski

Hasselt University, Diepenbeek, Belgium

Stuart G. Baker: sb16i@nih.gov

SUMMARY

Using multiple historical trials with surrogate and true endpoints, we consider various models to predict the effect of treatment on a true endpoint in a target trial in which only a surrogate endpoint is observed. This predicted result is computed using (1) a prediction model (mixture, linear, or principal stratification) estimated from historical trials and the surrogate endpoint of the target trial and (2) a random extrapolation error estimated from successively leaving out each trial among the historical trials. The method applies to either binary outcomes or survival to a particular time that is computed from censored survival data. We compute a 95% confidence interval for the predicted result and validate its coverage using simulation. To summarize the additional uncertainty from using a predicted instead of true result for the estimated treatment effect, we compute its multiplier of standard error. Software is available for download.

Keywords

Randomized trials; Reproducibility; Principal stratification

1. Introduction

In recent years, the medical and biostatistical literature has devoted considerable attention to surrogate versus true endpoints, where a true endpoint is the health outcome of interest and a surrogate endpoint is an outcome observed before the true endpoint that is used to make conclusions about the effect of intervention on the true endpoint (Weir and Walle, 2006, and Lassere, 2008). By using surrogate instead of true endpoints, clinicians can draw conclusions sooner, thereby potentially helping more future patients. However the use of surrogate endpoints to draw these conclusions involves the additional uncertainty associated with extrapolating estimates to an unobserved true endpoint.

To clearly and concisely discuss this extrapolation, we introduce the following terminology. A *target trial* is a randomized trial of the intervention of interest in which data are available on surrogate but not true endpoints. An *historical trial* is a previously conducted randomized trial with the same surrogate endpoint as in the target trial and the same true endpoint that would be observed with sufficiently long follow-up in the target trial, but typically comparing different interventions than in the target trial. A *true result* is the estimated effect of an intervention on the true endpoint in the target trial based on a comparison of the true endpoints from each arm of the target trial that would be observed with sufficiently long follow-up. A *surrogate result* is the estimated effect of an intervention on the surrogate endpoint in the target trial based on a comparison of surrogate endpoints from each arm of the target trial. A *predicted result* is the estimated effect of an intervention on the true endpoint in the target trial based on a comparison of predicted true endpoints from each arm of the target trial that are derived from the surrogate endpoints in the target trial and a model relating surrogate and true endpoints that is fit to data from the historical trials.

Analyzing trial data with surrogate endpoints requires an *extrapolation procedure*, which is a method to draw conclusions about the true result in the target trial based on either the surrogate or predicted result. When substituting a surrogate result for a true result in the target trial, the extrapolation procedure is testing a null hypothesis of no effect of treatment on the surrogate endpoint and drawing conclusions about a null hypothesis of no effect of treatment on true endpoint. When the surrogate endpoint is binary, this extrapolation procedure yields correct conclusions if the following criteria hold in the target trial: (i) the Prentice Criterion, namely that the true endpoint is conditionally independent of the randomization group given the surrogate endpoint, (ii) treatment is prognostic for the surrogate and true endpoints, and (iii) the surrogate endpoint is prognostic for the true endpoint (Prentice, 1989; Buyse and Molenberghs, 1998). When substituting a predicted result for the true result, an extrapolation procedure consists of computing a 95% confidence interval for the predicted result (Daniels and Hughes, 1997; Gail et al., 2002; Korn, Albert, and McShane, 2005).

Before using an extrapolation procedure to draw conclusions from a surrogate endpoint, it is important to compute a *validation measure*, which is a statistic based on historical trials to determine if the extrapolation procedure will likely yield correct conclusions. For a surrogate result, one validation measure is a test of the Prentice Criterion in historical trials. However, failure to reject the Prentice Criterion does not ensure it holds. Another validation measure for a surrogate result is the proportion of treatment effect explained in a historical trial (Freedman, Graubard, and Schatzkin, 1992). However, it is difficult to select an acceptable value of this measure. For a predicted result, a commonly used validation measure is R^2_{trial} , which equals one minus the ratio of the estimated variances of the true result adjusting, versus not adjusting, for the surrogate result in the historical trials (Buyse et al., 2000). However, selecting a threshold for R^2_{trial} that suffices for validation is difficult (Burzykowski and Buyse, 2006).

We consider surrogate and true endpoints involving probabilities of binary outcomes or probabilities of survival to a particular time computed from censored survival data. In this context, we make the following three contributions to the evaluation of an intervention in the target trial using the predicted result. First, we propose, as an extrapolation procedure, a 95% confidence interval for the predicted result which is derived from a prediction model and an estimated random extrapolation error. Second, we propose, as a validation measure, the coverage of the 95% confidence interval for the predicted result using a simulation based on the historical trials. Third, we propose, as a summary of the additional uncertainty when

using predicted instead of true results, a standard error multiplier based on the estimated variances of predicted and true results.

2. Overview of methodology

In this section we discuss the basic formulation, a key assumption, the extrapolation procedure, the validation measure, and the standard error multiplier.

2.1 Basic formulation

The basic formulation involves probabilities of binary surrogate and true outcomes that are later extended to survival data. Let subscript “ M ” denote the prediction model, which is discussed in detail in Section 3. The key idea is to write the predicted result for target trial i , denoted $(\text{predicted result})_{Mi}$, as the sum of a model-based predicted result, denoted $(\text{model result})_{Mi}$, and an estimated extrapolation error, denoted $\hat{\varepsilon}_{Mi}$,

$$(\text{predicted result})_{Mi} = (\text{model result})_{Mi} + \hat{\varepsilon}_{Mi}, \quad (1)$$

where $\hat{\varepsilon}_{Mi}$ has mean $\hat{\mu}_M$ and estimated variance $\hat{\sigma}_M^2$. Using methodology described in Section 3, we compute

$$(\text{model result})_{Mi} = \hat{\pi}_{Mi1} - \hat{\pi}_{Mi0}, \quad (2)$$

where $\hat{\pi}_{Mi z}$ is the estimated model-based probability of true endpoint in arm z of trial i . To compute the mean and variance of the estimated extrapolation error, we successively treat each historical trial j as a *left-out target trial* and the remaining trials as *left-in historical trials* and fit the following model to the extrapolation error, ε_{Mj} ,

$$(\text{true result for left – out trial})_j = (\text{model result for left – out trial})_{Mj} + \varepsilon_{Mj}, \quad (3)$$

where ε_{Mi} has expected value μ_M and variance σ_M^2 . The model-based predicted result for the left-out target trial is computed using the method described in Section 3. The true result for the left-out target trial is computed as

$$(\text{true result for left – out trial})_j = \hat{\pi}_{Tj1} - \hat{\pi}_{Tj0}, \quad (4)$$

where $\hat{\pi}_{Tj z}$ is the estimated probability of true endpoint in arm z of trial j . Based on Equation (3), the estimated extrapolation error for trial j is

$$\hat{\varepsilon}_{Mj} = (\text{true result for left – out trial})_j - (\text{model result for left – out trial})_{Mj}, \quad (5)$$

which has mean and estimated variance

$$\hat{\mu}_M = \sum_{j=1}^k \frac{\hat{\varepsilon}_{Mj}}{k} \text{ and } \hat{\sigma}_M^2 = \sum_{j=1}^k \frac{(\hat{\varepsilon}_{Mj} - \hat{\mu}_M)^2}{k-1}, \text{ respectively.} \quad (6)$$

2.2 Extrapolation procedure

The extrapolation procedure consists of computing a 95% confidence interval (CI) for predicted result $_{Mi}$ using a binomial estimate of the variance of (model result) $_{Mi}$ and the mean and estimated variance of the estimated extrapolation error,

$$CI_{Mi} = (\text{model result})_{Mi} + \hat{\mu}_M \pm 1.96 \sqrt{\widehat{\text{var}}(\text{model result})_{Mi} + \hat{\sigma}_M^2}, \quad (7)$$

where $\widehat{\text{var}}(\text{model result})_{Mi} = \sum_{z=0}^1 \hat{\pi}_{Miz} (1 - \hat{\pi}_{Miz}) / n_{iz}$ and n_{iz} is the number of persons in arm z of trial i .

2.3 Extrapolation assumption

To draw conclusions about the effect of treatment on the true endpoint in target trial i we assume that (predicted result) $_{Mi}$ in Equation (1) is an unbiased estimate of the effect of treatment on the unobserved true endpoint in target trial i . In the Discussion, we consider how this extrapolation assumption may be violated. We also assume that CI_{Mi} yields the correct coverage, which can not be fully checked by the validation measure.

2.4 Validation measure

The validation measure is the coverage of the 95% confidence interval for the predicted result as computed in a simulation of k historical trials and 1 target trial based on the data from the k historical trials. Ideally this coverage should be close to 95%. Each iteration of the simulation consists of the following steps.

- Step 1** Compute the mean and variance of the parameter estimates under the prediction model for each of k historical trials.
- Step 2** Randomly generate “population” parameters for k simulated historical trials and 1 simulated target trial based on the mean and variance in Step 1.
- Step 3** Compute “population” counts from the “population” parameters generated in Step 2.
- Step 4** Randomly generate “sample” counts from “population” counts in Step 3 using a multinomial model for the data from each trial arm.
- Step 5** From the “sample” counts in Step 4, use Equation (7) to compute the 95% confidence intervals for the predicted result in the simulated target trial.
- Step 6** From the “population” counts in Step 3, compute the “population” true result for the simulated target trial.
- Step 7** Set an indicator to 1 if the 95% confidence interval for the predicted result in Step 5 covers the “population” true result for the simulated target trial in Step 6, and set to 0 otherwise.

These steps are repeated for 1000 iterations. In Step 3, we use the sample sizes of the historical trials as sample sizes for the k simulated historical trials. For the simulated target trial, which corresponds to a yet unknown target trial, we use an average of the sample sizes over the k historical trials. If the target trial were known, its sample size would be used for the simulated target trial. In Step 4, to avoid numerical problems we added 0.1 to zero counts; adding 0.2 had little impact (not shown). The simulation coverage of each confidence interval of the predicted result is the fraction of iterations with indicator equal to 1 in Step 7. Details are given in Web Appendix A.

2.5 Standard error multiplier

Reasonably good coverage in the simulation is a necessary condition for the surrogate endpoint to be useful in computing predicted results. However it is not sufficient. There is a “price” for good coverage, namely wider confidence intervals that account for the variability of extrapolation. We quantify the increase width in the confidence intervals using the standard error multiplier, which is the average, over k historical trials, of the ratio of the standard error of the predicted result for each left-out trial to the standard error of the true result for each left-out trial,

$$(\text{standard error multiplier})_{M_j} = \frac{1}{k} \sum_{j=1}^k \frac{\sqrt{\widehat{\text{var}}(\text{model result})_{M_j} + \hat{\sigma}_{M_j}^2}}{\sqrt{\widehat{\text{var}}(\text{true result})_{M_j}}}, \quad (8)$$

where $\widehat{\text{var}}(\text{true result})_j = \sum_{z=0}^1 \hat{\pi}_{Tz} (1 - \hat{\pi}_{Tz}) / n_{Tz}$ and $\hat{\sigma}_{M_j}$ is computed using only the left-in historical trials. By modifying Equation (7), we also computed the 95% confidence interval for the predicted result of left-out target trial j .

3. Models for predicted results

We discuss the computation of $(\text{model result})_{M_i}$ for three types of models: mixture, linear, and principal stratification. Before discussing these models, we introduce basic parameters, an extension to survival data, weighted averages, and a method for labeling trial arms.

3.1 Basic parameters

Let $S = 0, 1$ denote the binary surrogate endpoint, and $T = 0, 1$ denote the binary true endpoint, where $S = 1$ and $T = 1$ corresponds to favorable outcomes. Let j index historical trials in a set H of historical trials, and let i index a target trial. The basic parameters are

$$\begin{aligned} \pi_{Sjz} &= pr(S=1|\text{historical trial } j, \text{ arm } z), \\ \pi_{Tjz} &= pr(T=1|\text{historical trial } j, \text{ arm } z), \\ \theta_{Tjzs} &= pr(T=1|S=s, \text{ historical trial } j, \text{ arm } z), \\ \pi_{Siz}^\circ &= pr(S=1|\text{target trial } i, \text{ arm } z). \end{aligned} \quad (9)$$

Let x_{jzst} denote the number of participants in historical trial j , arm z , with surrogate endpoint s and true endpoint t . Let x_{izs}° denote the number of participants in target trial i , arm z , with surrogate endpoint s . The estimates of the basic parameters are

$$\hat{\pi}_{Sjz} = \frac{x_{jz1+}}{x_{jz++}}, \quad \hat{\pi}_{Tjz} = \frac{x_{jz+1}}{x_{jz++}}, \quad \hat{\theta}_{Tjzs} = \frac{x_{jzsl}}{x_{jzsl+}}, \quad \hat{\pi}_{Siz}^\circ = \frac{x_{izl}^\circ}{x_{iz+}^\circ}, \quad (10)$$

where the subscript “+” denotes summation over the indicated index.

3.2 Extension to survival data

With survival data, we compute the basic parameters by first transforming the survival data into counts for binary outcomes. Let $S = 0$ if an unfavorable surrogate event (e.g. cancer recurrence) occurs before time τ_S , and 1 otherwise. Let $T = 0$ if an unfavorable true event (e.g. death) occurs before time τ_T , and 1 otherwise. Let π_{Sjz} and θ_{Tjzs} denote Kaplan-Meier estimates (here with discrete times) of π_{Sjz} and θ_{Tjzs} , respectively. For specifying s in θ_{Tjzs}

there are two considerations. First the surrogate event may be monitored separately from the true event so the surrogate event can be censored when the true event is observed. Second, if a true event occurs before a given time, the surrogate event must have occurred before that time even if the surrogate event is censored before that time. With these considerations, we define $s = 0$ if either the surrogate or the true event is observed before τ_S , we define $s = 1$ if the surrogate event is observed or is censored after τ_S , and we discard the typically very small subset censored for both surrogate and true events before τ_S . Letting n_{jz} denote the number of subjects in trial j and arm z , the cell counts derived from the individual-level survival data are

$$\begin{aligned}\tilde{x}_{jz00} &= n_{jz}(1 - \tilde{\pi}_{Sjz})(1 - \tilde{\theta}_{T|jz0}), & \tilde{x}_{jz10} &= n_{jz}(1 - \tilde{\pi}_{Sjz})\tilde{\theta}_{T|jz0}, \\ \tilde{x}_{jz10} &= n_{jz}\tilde{\pi}_{Sjz}(1 - \tilde{\theta}_{T|jz1}), & \tilde{x}_{jz11} &= n_{jz}\tilde{\pi}_{Sjz}\tilde{\theta}_{T|jz1}.\end{aligned}\quad (11)$$

As shown in Appendix A, substituting $\tilde{x}_{jzst}$ from Equation (11) into Equation (10) yields the same estimates and similar variances to those obtained directly from individual-level survival data. The counts in Equation (11) can be made publicly available (Web Tables 1, 2, 3) even when individual-level survival data cannot be made available due to confidentiality issues. Public availability of data is important so that other researchers can reproduce the results.

3.3 Using a weighted average

When computing model-based predicted results, we use a weighted average of model-based predicted results from each historical trial, where the weights are proportional to sample size. We use the notation $avg_{j \in H}()$ for the weighted average over the set H of historical trials.

3.4 Labeling of trial arms

Sometimes is not clear how to label treatment arms as $Z = 0$ or $Z = 1$, particularly when a treatment is a control in one trial and experimental in another trial. To make labeling objective, we assign $Z = 1$ to the arm of the trial in which the estimate of $pr(S = 1)$ is largest and $Z = 0$ to the arm of the trial in which the estimate of $pr(S = 1)$ is smallest (Baker, 2008). Recall that $pr(S = 1)$ is either the probability of survival or the probability of the more favorable outcome.

3.5 Mixture model

The mixture model (Baker, 2008) is based on the mathematical identity

$$\pi_{Tjz} = \theta_{T|jz0}(1 - \pi_{Sjz}) + \theta_{T|jz1}\pi_{Sjz}. \quad (12)$$

Equation (12) motivates the following model (MIX1) for the predicted probability of the true endpoint in target trial i based on an estimate of $\theta_{T|jz}$ from historical trial j and an estimate of π_{Siz}° from target trial i ,

$$\hat{\pi}_{\text{MIX1}ijz} = \hat{\theta}_{T|jz0}(1 - \hat{\pi}_{Siz}^\circ) + \theta_{T|jz1}\hat{\pi}_{Siz}^\circ. \quad (13)$$

No additional assumptions are needed for estimation. The model-based predicted result for target trial i is the average difference in predicted probabilities,

$$\begin{aligned} (\text{model result})_{\text{MIX1}i} &= \text{avg}_{j \in H} (\widehat{\pi}_{\text{MIX1}i1j} - \widehat{\pi}_{\text{MIX1}i0j}) \\ &= \widehat{\pi}_{\text{MIX1}i1} - \widehat{\pi}_{\text{MIX1}i0}, \text{ where } \widehat{\pi}_{\text{MIX1}iz} = \text{avg}_{j \in H} (\widehat{\pi}_{\text{MIX1}izj}). \end{aligned} \quad (14)$$

The additional assumption of the Prentice Criterion, namely that $\hat{\theta}_{Tjcs} \equiv \hat{\theta}_{Tjs}$, leads to mixture model MIX2,

$$\widehat{\pi}_{\text{MIX2}izj} = \widehat{\theta}_{Tj0} (1 - \widehat{\pi}_{Siz}^{\circ}) + \widehat{\theta}_{Tj1} \widehat{\pi}_{Siz}^{\circ}. \quad (15)$$

Based on Equation (15), the model-based predicted result under MIX2 is

$$\begin{aligned} (\text{model result})_{\text{MIX2}i} &= \text{avg}_{j \in H} (\widehat{\pi}_{\text{MIX2}i1j} - \widehat{\pi}_{\text{MIX2}i0j}) \\ &= \widehat{\pi}_{\text{MIX2}i1} - \widehat{\pi}_{\text{MIX2}i0}, \text{ where } \widehat{\pi}_{\text{MIX2}iz} = \text{avg}_{j \in H} (\widehat{\pi}_{\text{MIX2}izj}), \\ &= (\widehat{\pi}_{S11}^{\circ} - \widehat{\pi}_{S10}^{\circ}) \text{avg}_{j \in H} (\widehat{\theta}_{Tj1} - \widehat{\theta}_{Tj0}). \end{aligned} \quad (16)$$

which is proportional to the surrogate result, namely $\widehat{\pi}_{S11}^{\circ} - \widehat{\pi}_{S10}^{\circ}$.

3.6 Linear model

A random effects linear model for predicted results can be derived by assuming a joint multivariate normal distribution for parameters modeling surrogate and true endpoints in the historical trials (Buyse et al., 2000; Gail et al., 2000). The following model (LIN1) is a variation of this model involving binary surrogate and true endpoints,

$$\begin{aligned} \pi_j = (\pi_{S10}, \pi_{T10}, \pi_{S11}, \pi_{T11}) &\sim \text{Normal}(\pi, \Sigma), \text{ where} \\ \pi = (\pi_{S0}, \pi_{T0}, \pi_{S1}, \pi_{T1}) \text{ and } \Sigma &= \begin{pmatrix} d_{S0S0} & d_{S0T0} & d_{S0S1} & d_{S0T1} \\ d_{S0T0} & d_{T0T0} & d_{S1T0} & d_{T0T1} \\ d_{S0S1} & d_{S1T0} & d_{S1S1} & d_{S1T1} \\ d_{S0T1} & d_{T0T1} & d_{S1T1} & d_{T1T1} \end{pmatrix}. \end{aligned} \quad (17)$$

A standard formula for a conditional multivariate normal distribution transforms Equation (17) into the multivariate linear regression,

$$\begin{aligned} E\left(\begin{pmatrix} \pi_{Tj0} \\ \pi_{Tj1} \end{pmatrix} \middle| \begin{pmatrix} \pi_{Sj0} \\ \pi_{Sj1} \end{pmatrix}\right) &= \begin{pmatrix} \pi_{T0} \\ \pi_{T1} \end{pmatrix} + M \left\{ \begin{pmatrix} \pi_{Sj0} \\ \pi_{Sj1} \end{pmatrix} - \begin{pmatrix} \pi_{S0} \\ \pi_{S1} \end{pmatrix} \right\}, \text{ where} \\ M &\equiv \begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix} = \begin{pmatrix} d_{S0T0} & d_{S0T1} \\ d_{S1T0} & d_{S1T1} \end{pmatrix} \begin{pmatrix} d_{S0S0} & d_{S0S1} \\ d_{S0S1} & d_{S1S1} \end{pmatrix}^{-1}. \end{aligned} \quad (18)$$

The model-based predicted result under LIN1 for trial i is

$$\begin{aligned} (\text{model result})_{\text{LIN1}i} &= \widehat{\pi}_{\text{LIN1}i1} - \widehat{\pi}_{\text{LIN1}i0}, \text{ where } \widehat{\pi}_{\text{LIN1}iz} = \text{avg}_{j \in H} (\widehat{\pi}_{\text{LIN1}izj}), \\ \begin{pmatrix} \widehat{\pi}_{\text{LIN1}i0j} \\ \widehat{\pi}_{\text{LIN1}i1j} \end{pmatrix} &= \begin{pmatrix} \widehat{\pi}_{T0j} \\ \widehat{\pi}_{T1j} \end{pmatrix} + \widehat{M}_{\text{LIN1}} \left\{ \begin{pmatrix} \widehat{\pi}_{Sj0}^{\circ} \\ \widehat{\pi}_{Sj1}^{\circ} \end{pmatrix} - \begin{pmatrix} \widehat{\pi}_{S0j}^{\circ} \\ \widehat{\pi}_{S1j}^{\circ} \end{pmatrix} \right\}, \end{aligned} \quad (19)$$

and $\widehat{M}_{\text{LIN1}}$ is computed from an estimate of Σ (Web Appendix B).

The model LIN2 is similar to LIN1 but with a logit transformation for the model-based predicted result,

$$\begin{aligned}
 (\text{model result})_{\text{LIN2}i} &= \widehat{\pi}_{\text{LIN2}i1} - \widehat{\pi}_{\text{LIN2}i0}, \text{ where } \widehat{\pi}_{\text{LIN2}i2} = \text{avg}_{j \in H} \left\{ \frac{\exp(\widehat{\varphi}_{\text{LIN2}ij})}{1 + \exp(\widehat{\varphi}_{\text{LIN2}ij})} \right\}, \\
 \begin{pmatrix} \widehat{\varphi}_{\text{LIN2}i0j} \\ \widehat{\varphi}_{\text{LIN2}i1j} \end{pmatrix} &= \begin{pmatrix} \text{logit}(\widehat{\pi}_{T0j}) \\ \text{logit}(\widehat{\pi}_{T1j}) \end{pmatrix} + \widehat{M}_{\text{LIN2}} \left\{ \begin{pmatrix} \text{logit}(\widehat{\pi}_{S0}^{\circ}) \\ \text{logit}(\widehat{\pi}_{S1}^{\circ}) \end{pmatrix} - \begin{pmatrix} \text{logit}(\widehat{\pi}_{S0j}) \\ \text{logit}(\widehat{\pi}_{S1j}) \end{pmatrix} \right\},
 \end{aligned} \tag{20}$$

where $\widehat{M}_{\text{LIN2}}$ is computed using the delta method (Web Appendix C).

The model LIN3 is a special case of LIN1 under the following assumptions: (i) $d_{S0S0} = d_{S1S1}$, (ii) $d_{S0S1} = d_{S0T1} = d_{S1T0} = 0$, which is reasonable if one arm does not influence another arm, and (iii) $d_{S0T0} = d_{S1T1}$, which is similar to the Prentice Criterion. As derived in Appendix B, the model-based predicted result under LIN3 has the form,

$$(\text{model result})_{\text{LIN3}i} = \widehat{\alpha}_{\text{LIN3}} + \widehat{\beta}_{\text{LIN3}} (\text{surrogate result})_i. \tag{21}$$

We compute $\widehat{\alpha}_{\text{LIN3}}$ and $\widehat{\beta}_{\text{LIN3}}$ using weighted least squares with weights proportional to the sample size. To put the model-based predicted result for LIN3 into our standard form of a difference in probabilities, we rewrite Equation (21) as

$$\begin{aligned}
 (\text{model result})_{\text{LIN3}i} &= \widehat{\pi}_{\text{LIN3}i1} - \widehat{\pi}_{\text{LIN3}i0}, \text{ where} \\
 \widehat{\pi}_{\text{LIN3}i1} &= \widehat{\alpha}_{\text{LIN3}} / 2 + \widehat{\beta}_{\text{LIN3}} \widehat{\pi}_{S1}^{\circ}, \\
 \widehat{\pi}_{\text{LIN3}i0} &= -\widehat{\alpha}_{\text{LIN3}} / 2 + \widehat{\beta}_{\text{LIN3}} \widehat{\pi}_{S0}^{\circ},
 \end{aligned} \tag{22}$$

arbitrarily splitting $\widehat{\alpha}_{\text{LIN3}}$ in half.

Model LIN4 is a special case of LIN3 that computes $\widehat{\beta}_{\text{LIN4}}$ from a weighted least squares model with a zero intercept,

$$(\text{model result})_{\text{LIN4}i} = \widehat{\pi}_{\text{LIN4}i1}, \text{ where } \widehat{\pi}_{\text{LIN4}i2} = \widehat{\beta}_{\text{LIN4}} \widehat{\pi}_{S2}^{\circ}. \tag{23}$$

3.7 Principal stratification model

Permutt and Hebel (1989), Baker and Lindeman (1994), Imbens and Angrist (1994) followed by Angrist, Imbens, and Rubin (1996), and Cuzick, Edwards, and Segnan (1997) independently proposed a potential outcomes model involving a group 0 in which all subjects would ideally receive treatment T0 and a group 1 in which all subjects would ideally receive treatment T1 with the following two key features. First, it involved four potential outcome categories: (i) receive treatment T1 regardless of group, (ii) receive treatment T0 if in group 0 and T1 if in group 1, (iii) receive treatment T1 if in group 0 and T0 if in group 1, and (iv) receive treatment T0 regardless of group. Second, under plausible assumptions of no person in category (iii) and the effect of treatment not depending on group in categories (i) and (iv), the model yields an unbiased estimate of the effect of receipt of treatment in category (ii), which is then generalized to all subjects.

Frangakis and Rubin (2002) extended the four potential outcomes categories to any post-randomization variable, including a binary surrogate endpoint, and called the four categories principal strata. An appealing feature of this model is that the principal strata are baseline variables. Frangakis and Rubin (2002) proposed using principal stratification to predict a true endpoint using data from surrogate endpoints in a target trial and a single historical trial, but did not discuss estimation. We extend their proposal to multiple historical trials and

discuss estimation. Let $S_{PS} = (a, b)$ denote a principal stratum for the surrogate endpoint, where a is the level of a binary surrogate endpoint if a subject were randomized to the control arm $z = 0$, and b is the level of a binary surrogate endpoint if a subject were randomized to the experimental arm $z = 1$. The parameters of the model are

$$\begin{aligned}\pi_{Sj(ab)} &= pr(S_{PS} = (a, b) | \text{historical trial } j), \\ \theta_{T|jz(ab)} &= pr(T=1 | S_{PS} = (a, b), \text{historical trial } j, \text{arm } z), \\ \pi_{Si(ab)}^\circ &= pr(S_{PS} = (a, b) | \text{target trial } i).\end{aligned}\quad (24)$$

Without assumptions, the general principal stratification model (PSM),

$$\pi_{Tjz} = \sum_{a=0}^1 \sum_{b=0}^1 \theta_{T|jz(ab)} \pi_{Sj(ab)}, \quad (25)$$

is not identifiable because the number of parameters exceeds the number of independent cell counts. To obtain an identifiable model, we make the following assumptions which are analogous to the aforementioned assumptions in potential outcomes model involving treatments T0 and T1.

Assumption PSM-1—No subject would have a surrogate endpoint at level $s = 1$ if randomized to arm $z = 0$ and a surrogate endpoint at level $s = 0$ if randomized to arm $z = 1$, implying $\pi_{Sj(10)} = 0$ and $\pi_{Si(10)}^\circ = 0$.

Assumption PSM-2—The probability of true endpoint does not depend on arm z among subjects whose level of surrogate endpoint s would remain unchanged with a different arm z , implying $\theta_{T|j0(aa)} = \theta_{T|j1(aa)} \equiv \theta_{T|j(aa)}$, for $a = 0, 1$.

Assumption PSM-1 is plausible because of the labeling scheme that defines $z = 1$ as the arm with the highest estimated probability that $s = 1$. *Assumption PSM-2* is plausible if treatment affects only recurrence of the primary cancer; however it could be violated if treatment affects microscopic metastases that could lead to a secondary tumor. Incorporating these assumptions gives the principal stratification model

$$\pi_{Tjz} = \sum_{a=0}^1 \theta_{T|j(aa)} \pi_{Sj(aa)} + \theta_{T|jz(01)} \pi_{Sj(01)}. \quad (26)$$

Based on Equation (26), the model-based predicted result under PSM is

$$\begin{aligned}(\text{model result})_{PSMi} &= \widehat{\pi}_{PSMi1} - \widehat{\pi}_{PSMi0}, \text{ where} \\ \widehat{\pi}_{PSMi1} &= \sum_{a=0}^1 \text{avg}_{j \in H} \{\widehat{\theta}_{T|j(aa)}\} \widehat{\pi}_{Si(aa)}^\circ + \text{avg}_{j \in H} \{\widehat{\theta}_{T|iz(01)}\} \widehat{\pi}_{Si(01)}^\circ.\end{aligned}\quad (27)$$

Maximum likelihood estimates of parameters from the historical trials are computed as a perfect fit estimate, if admissible, or otherwise via an EM algorithm (Baker, 2011) with $\widehat{\theta}_{T|iz(01)}$ restricted between 0.01 and 0.99. For the target trial, the perfect fit estimates are $\widehat{\pi}_{Si(00)}^\circ = 1 - \widehat{\pi}_{Si1}^\circ$, $\widehat{\pi}_{Si(11)}^\circ = \widehat{\pi}_{Si0}^\circ$ and $\widehat{\pi}_{Si(01)}^\circ = 1 - \widehat{\pi}_{Si(00)}^\circ - \widehat{\pi}_{Si(11)}^\circ$ which implies that the model-based predicted result is proportional to the surrogate result,

$$(\text{model result})_{\text{PSM}_i} = \text{avg}_{j \in H} \{ \hat{\theta}_{T|jz(01)} \} (\hat{\pi}_{S1}^{\circ} - \hat{\pi}_{S0}^{\circ}). \quad (28)$$

If the perfect fit estimate for $\pi_{Si(01)}^{\circ}$ is less than 0, we instead set $\hat{\pi}_{Si(11)}^{\circ} = 1 - \hat{\pi}_{Si(00)}^{\circ}$ and $\hat{\pi}_{Si(01)}^{\circ} = 0$.

4. Application

4.1 Data Sets

We applied our method to three data sets of historical trials (Web Tables 1, 2, and 3) and an artificial data set that combined subsets of these three data sets. Before looking at the data, we specified the survival times for surrogate and true events, τ_S and τ_T , that are needed for classification into surrogate and true endpoints.

Data Set 1 consists of 10 randomized trials for early colon cancer (Sargent et al., 2005) with $S = 0$ if cancer recurred before 3 years, and $S = 1$ otherwise. Also $T = 0$ if overall mortality occurred before 5 years, and $T = 1$ otherwise. Eight trials had two arms, one trial had three arms, and one trial had four arms. All of the trials compared different sets of treatments, with some treatments given in multiple trials. To estimate $\theta_{T|jzS}$ we discarded the 1% of subjects who were censored for both surrogate and true events before 3 years.

Data Set 2 consists of 10 randomized trials for advanced colorectal cancer (Meta-Analysis Group in Cancer, 2004; Buyse et al., 2007) with $S = 0$ if cancer progressed before 6 months, and $S = 1$ otherwise. Also $T = 0$ if overall mortality occurred before 12 months and $T = 1$ otherwise. All of the trials had two arms. In 7 of the ten trials, the intervention was FU + LV versus FU, with FU+LV labeled as the control arm in 6 of those trials and the experimental arm in the other trial. In the remaining 3 trials, the intervention was FU+LV versus RA. This is not the ideal situation in which all the interventions differ. To estimate $\theta_{T|jzS}$ we discarded the 0.2% of subjects who were censored for both surrogate and true events before 6 months.

Data Set 3 consists of 27 randomized trials for advanced colorectal cancer (Burzykowski et al., 2004) with $S = 0$ if the cancer was stable or progressive at 3–6 months, and $S = 1$ if complete or partial response of the tumor occurred at 3–6 months. Also $T = 0$ if overall mortality occurred before 12 months, and 0 otherwise. All of the trials had two arms. A few trials in Data Set 3 were the same as in Data Set 2. Because the surrogate endpoint was binary, all subjects contributed to the estimation of $\theta_{T|jzS}$.

Data Set 4 is an artificial data set of 10 randomized trials that consists of data from the first three trials in Data Sets 1 and 2 and the first four trials in Data Set 3. The surrogate and true endpoints differ in each of three data sets. Also the relationship of surrogate and true endpoints likely differs substantially among the data sets. Therefore a prediction rule based on a combination of historical trials from three data sets will not likely perform well for a target trial from only one of the data sets. As will be discussed, the standard error multipliers for Data Set 4 were larger than for the other data sets, which is important for showing that the methodology can identify a poor surrogate endpoint.

4.2 Confidence intervals for predicted result

Applying Equation (7) to the data sets, we computed 95% confidence intervals for predicted result for left-out trials and plotted these confidence intervals along with the 95% confidence intervals for the true result for left-out trials (Figure 1 and Web Figures 1–4). To illustrate the computations, we present estimates for model LIN4 in Data Set 1 (Table 1). Because the 95% confidence intervals for the predicted results include the estimated variance of the estimated extrapolation error, they are wider than the corresponding 95% confidence

intervals for the true results. Due to random variability, the confidence interval for the true result for the left-out trial is not a perfect gold standard for the confidence interval for the predicted result for the left-out trial. The validation measure provides better information on the appropriateness of the confidence interval for the predicted result.

4.3 Validation measure

We computed the validation measure of coverages of 95% confidence intervals for each model using separate simulations under mixture, linear, and principal stratification models derived from the data (Table 2). For Data Sets 1, 2, and 3, most of the coverages were close to the target of 95%, and for the artificial Data Set 4 the coverages were slightly worse. In our framework, this validation measure does not distinguish between good and poor surrogate endpoints because reasonable coverage can be obtained with a poor surrogate endpoint due to its compensating large standard error multiplier. Instead the validation measure checks a necessary condition for drawing conclusions from the extrapolation procedure.

4.4 Standard error multiplier

We evaluated the quality of the surrogate endpoint using the standard error multiplier (Table 3). The smallest standard error multipliers generally corresponded to models MIX1, LIN3, LIN4, and PSM. Not surprisingly, the standard error multipliers were largest for Data Set 4, which was created to reflect a poor surrogate endpoint. It is possible to decrease the standard error multiplier by increasing the sample size of a target trial. However, for our data sets, an increase in sample size of 20% in the left-out target trials yielded only a small reduction in the standard error multiplier (Table 3) because the extrapolation error accounts for a large fraction of the variance.

5. Discussion

With survival data, our focus is on probabilities of survival at specific times, as opposed to median survival times or hazard functions. Mathematically, an advantage of using the probability of survival at a specific time rather than the median survival time is the ability to use a binomial distribution to compute a variance for the model-based predicted result. However, the probability of survival at specific time is also of great interest. In oncology, cancer survival rates, which are the percentage of people who survive a certain type of cancer for a specific amount of time, are often used to inform patients (Mayo Clinic staff, 2009). For the true endpoint, the specified time for survival should be the survival time used in standard evaluations of treatment for the disease. For a surrogate endpoint, the specified time for survival could be based on expert judgment that weighs the benefit of reaching conclusions sooner against the loss of information. Alternatively, for surrogate endpoints the specified time for survival could be chosen to give an overall (combining data over trial arms) survival probability for the surrogate event similar to the overall survival probability for the true event at its specified time.

The labeling of trial arms deserves additional discussion. One way to reduce sensitivity to labeling is by modeling the predicted results as proportional to the surrogate results; however labeling can still affect estimation for the extrapolation error. Our labeling method involving the highest and lowest probabilities of surrogate endpoints has the desirable feature of not depending on a standard of care that can change over time. Nevertheless, for a sensitivity analysis we also computed standard error multipliers using labels based on the standard of care at the time of the trial (Web Table 4). The standard error multipliers under the two labeling methods were similar, particularly for Data Sets 1, 2, and 3. For a trial with three arms, reporting two pair-wise comparisons may not be desirable because one arm

would contribute twice to the calculations. Our labeling method avoids this problem by selecting two arms from a trial with more than two arms.

The standard error multipliers provide information on the usefulness of the predicted result and hence on the usefulness of the underlying surrogate endpoint. Under most of the models, the predicted results in Data Sets 1 and 3 had smaller standard error multipliers than the predicted results in Data Set 2, indicating better surrogacy in Data Sets 1 and 3 than in Data Set 2. This differs from previous findings of acceptable surrogate endpoints in Data Sets 1 and 2 and a poor surrogate endpoint in Data Set 3 (Burzykowski, et al 2004; Buyse et al 2008). A possible explanation is that the previous methods estimated the effect of treatment on a hazard function for the true endpoint while our method estimates the effect of treatment on a difference in probabilities of a true endpoint, which may be more closely related to the difference in probabilities of the binary surrogate endpoint.

A comparison of the standard error multipliers by model is also of interest. It is not clear why the simplest linear model LIN4 performed the best of the linear models. One possible explanation is that the model-based predicted result under LIN4 is proportional to the surrogate result which can reduce bias from incorrect labeling if the mean of the extrapolated error is small. Another possible explanation is less overfitting than with a more complicated linear model. In contrast, the mixture model MIX1 performed substantially better than the simpler model MIX2 in which the predicted result is proportional to the surrogate result. Apparently MIX2 is too overly restrictive. The principal stratification model PSM performed similarly to MIX1, perhaps because both represent saturated models. Among the best fitting models MIX1 and LIN4 are the simplest to compute and hence recommended based on this investigation.

The key to drawing correct conclusions from the predicted result is the appropriateness of the extrapolation assumption that the predicted result derived from historical trials correctly applies to the target trial. Unfortunately this assumption is not testable unless there is follow-up to observe the true endpoints, in which case the use of surrogate endpoints would be superfluous. To buttress the extrapolation assumption, investigators should strive to satisfy the following desiderata:

1. all trials should apply to the same disease,
2. the assigned treatments in all trials should affect surrogate and true endpoints via the same general mechanism,
3. any intervention administered after the surrogate endpoint and before the true endpoint should be similar among all trials.

To further strengthen conclusions, investigators can use the scheme of Lassere (2008) that also considers the quality of historical trials, the clinical characteristics of the true endpoint, and evidence from animal experiments and epidemiological studies. Also investigators need to decide if there is sufficient time to observe any detrimental side effects in the target trial when there is no follow-up to observe the true endpoint (Ellenberg 1993).

Interested readers can reproduce results or evaluate other data sets using software written in Mathematica 8.0 (Wolfram Research, 2008), which is available for download at <http://prevention.cancer.gov/programs-resources/groups/b/software/endpoint>.

6. Supplementary Materials

Web Appendices and Tables referenced in Sections 2, 3, and 4 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Cancer Institute. The authors are grateful to the MAGIC (Meta-Analysis Group in Cancer) and ACCENT (Adjuvant Colon Cancer Endpoints) collaborators, listed in Burzykowski (2008), for providing the data. The authors thank the reviewers for helpful comments.

References

- Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. 1996; 92:444–455.
- 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]
- Baker SG. A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. *Biostatistics*. 2006; 7:57–70.
- Baker SG. Two simple approaches for validating a binary surrogate endpoint using data from multiple trials. *Statistical Methods in Medical Research*. 2008; 17:505–514. [PubMed: 18285436]
- Baker SG. Estimation and inference for the causal effect of receiving treatment on a multinomial outcome: an alternative approach. *Biometrics*. 2011; 67:319–325. [PubMed: 20560933]
- Burzykowski T, Buyse M. Surrogate threshold effect: An alternative for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*. 2006; 5:173–186. [PubMed: 17080751]
- Burzykowski T, Molenberghs G, Buyse M. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society Series A*. 2004; 167:103–124.
- Burzykowski T, Buyse M, Yothers G, Sakamoto J, Sargent D. Exploring and validating surrogate endpoints in colorectal cancer. *Lifetime Data Analysis*. 2008; 14:54–64. [PubMed: 18205045]
- Buyse M, Burzykowski T, Carroll K, Michiels S, Sargent D, Miller LL, Elfring GL, Pignon JP, Piedbois P. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*. 2007; 25:5218–5224. [PubMed: 18024867]
- Buyse M, Burzykowski T, Michiels S, Sargent D, Carroll K. Individual- and trial- level surrogacy in colorectal cancer. *Statistical Methods in Medical Research*. 2008; 17:467–475. [PubMed: 18285439]
- Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998; 54:1014–1029. [PubMed: 9840970]
- Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000; 1:49–67. [PubMed: 12933525]
- Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*. 1997; 16:1017–1029. [PubMed: 9160496]
- Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*. 1997; 16:1965–1982. [PubMed: 9304767]
- Ellenberg SS. Surrogate endpoints. *British Journal of Cancer*. 1993; 68:457–459. [PubMed: 8353034]
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
- Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine*. 1992; 11:167–178. [PubMed: 1579756]
- Gail MH, Pfeiffer R, Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics*. 2000; 3:231–246. [PubMed: 12933506]
- Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica*. 1994; 62:467–475.

- Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*. 2010; 29:1282–1297. [PubMed: 19408255]
- Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed model. *Statistics in Medicine*. 2005; 24:63–182.
- 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]
- Mayo Clinic staff. Cancer survival rate: What it means for your prognosis. 2009. <http://www.mayoclinic.com/health/cancer/CA00049>
- Meta-Analysis Group in Cancer. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: an updated meta-analysis. *Journal of Clinical Oncology*. 2004; 22:3766–3775. [PubMed: 15365073]
- Permutt T, Hebel R. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*. 1989; 45:619–622. [PubMed: 2669989]
- Prentice RL. Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine*. 1989; 8:431–430.
- Sargent DJ, Wieand S, Haller DG, Gray R, Benedetti J, Buyse M, Labianca R, Seitz JF, Callaghan CJO, Francini G, Grothey A, O'Connell M, Catalano PJ, Blanke CD, Kerr D, Green E, Wolmark N, Andre T, Goldberg RM, De Gramont A. Disease-free survival (DFS) vs. overall survival (OS) as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology*. 2005; 23:8664–8670. [PubMed: 16260700]
- 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]
- Wolfram Research, Inc. Mathematica, Version 8.0. Champaign, IL: 2010.

Appendix A

Without loss of generality, consider two cells representing survival or not. First we show that estimates of survival are the same when computed from expected counts or from the individual-level survival data. Let S_y denote the estimated probability of surviving y time intervals based on individual-level data, so that $S_y = \prod_{u=1}^y (1 - h_u)$, where h_u is the estimated hazard in time interval u . Let n denote the total number of patients. The two estimated cell counts are

$$x_0 = n(1 - S_y) \text{ and } x_1 = n S_y, \quad (\text{A.1})$$

Let S_{BIN} denote the estimate survival probability based on the estimated counts,

$$S_{BIN} = x_1 / x_+. \quad (\text{A.2})$$

Substituting (A.1) into (A.2) gives $S_{BIN} = S_y$.

Second we show that estimates of the variances of the estimated survival probability are similar when computed from expected counts or from the individual-level survival data. The estimated variance of S_y based on the individual survival data is

$$\widehat{var}_{SU RV}(S_y) = S_y^2 \sum_{u=1}^y \frac{h_u}{(1-h_u)r_u}, \quad \text{(A. 3)}$$

where r_u is the number at risk in time interval u . Approximating r_u by the expected number at risk in time interval u namely $r_u = n S_{u-1}$, gives

$$\widehat{var}_{SU RV}(S_y) = \frac{S_y}{n} [h_k + h_{k-1}(1-h_k) + h_{k-2}(1-h_{k-1})(1-h_k) + \dots]. \quad \text{(A. 4)}$$

Because the quantity in brackets in (A.4) is the estimated probability of failure in y intervals, it equals $1 - S_y$, so that

$$\widehat{var}_{SU RV}(S_y) = S_y(1 - S_y)/x_+. \quad \text{(A. 5)}$$

The estimated variance of S_y based on the estimated counts is

$$\widehat{var}_{BIN}(S_{BIN}) = S_{BIN}(1 - S_{BIN})/x_+. \quad \text{(A. 6)}$$

Substituting (A.2) into (A.6) gives $\widehat{var}_{BIN}(S_{BIN}) = \widehat{var}_{SR RV}(S_y)$.

Appendix B

Equation (19) can be written as

$$\begin{aligned} (\text{model result})_{LIN1i} &= (\bar{\pi}_{T1j} - \bar{\pi}_{T0j}) + (m_{11} - m_{01})(\bar{\pi}_{S1i}^\circ - \bar{\pi}_{S1j}) + (m_{10} - m_{00})(\bar{\pi}_{S0i}^\circ - \bar{\pi}_{S10}) \\ &= (\bar{\pi}_{T1j} - \bar{\pi}_{T0j}) + (m_{11} - m_{01})\{(\bar{\pi}_{S1i}^\circ - \bar{\pi}_{S1j}) - (\bar{\pi}_{S0i}^\circ - \bar{\pi}_{S10})\} + \{(m_{10} - m_{00}) + (m_{11} - m_{01})\}(\bar{\pi}_{S0i}^\circ - \bar{\pi}_{S10}) \end{aligned}$$

Under the assumptions for LIN3, $\{(m_{10} - m_{00}) + (m_{11} - m_{01})\} = 0$, giving the general form of Equation (21) in which the model-based predicted result for LIN3 equals a constant plus a multiple of the surrogate result.

Figure 1.
Point estimates and 95% confidence intervals for true and predicted results in Data Set 1.

Table 1

Example of calculations: Model LIN4 in Data set 1

left-out target trial j	true result ¹		model-based predicted result ²		estimated extrapolation error ³			predicted result ⁴	
	estimate	std err	estimate	std err	mean	std. dev	estimate	std err	
1	0.036	0.036	0.048	0.036	0.004	0.028	0.052	0.045	
2	0.089	0.062	0.117	0.062	0.006	0.026	0.123	0.067	
3	0.073	0.031	0.120	0.030	0.008	0.022	0.127	0.037	
4	0.063	0.029	0.029	0.031	−0.001	0.026	0.027	0.040	
5	0.093	0.028	0.063	0.030	−0.001	0.026	0.063	0.040	
6.	0.102	0.045	0.067	0.046	−0.001	0.026	0.066	0.053	
7	0.037	0.024	0.034	0.025	0.002	0.028	0.037	0.038	
8	0.056	0.048	0.052	0.049	0.002	0.028	0.054	0.057	
9	0.008	0.018	0.012	0.019	0.003	0.028	0.015	0.034	
10	0.023	0.031	0.014	0.032	0.002	0.028	0.016	0.043	

¹ Columns corresponds to $(\text{true result})_j$ and $\sqrt{\widehat{\text{var}}(\text{true result})_j}$.

² Columns corresponds to $(\text{model result})_{\text{LIN4}j}$ and $\sqrt{\widehat{\text{var}}(\text{model result})_{\text{LIN4}j}}$.

³ Columns corresponds to $\hat{\mu}_{\text{LIN4}j}$ and $\hat{\sigma}_{\text{LIN4}j}$.

⁴ Columns corresponds to $(\text{predicted result})_{\text{LIN4}j} = (\text{model result})_{\text{LIN4}j} + \hat{\mu}_{\text{LIN4}j}$ and $\sqrt{\widehat{\text{var}}(\text{model result})_{\text{LIN4}j} + \hat{\sigma}_{\text{LIN4}j}^2}$.

The standard error multiplier is the average of the ratio of the standard error of the predicted result (column 9) to the standard error of the true result (column 3).

Table 2
Validation measure: coverage of 95% confidence interval based on 1000 simulations

Data Set	Data generation model	Data analysis model							
		MIX1	MIX2	LIN1	LIN2	LIN3	LIN4	PSM	PSM
1	MIX	0.980	0.948	0.962	0.967	0.978	0.967	0.967	0.968
	LIN	0.993	0.992	0.980	0.981	0.985	0.979	0.998	0.998
	PSM	0.972	0.985	0.956	0.956	0.965	0.950	0.977	0.977
2	MIX	0.947	0.946	0.933	0.920	0.937	0.933	0.939	0.939
	LIN	0.973	0.977	0.969	0.975	0.958	0.960	0.977	0.977
	PSM	0.963	0.961	0.953	0.951	0.939	0.933	0.947	0.947
3	MIX	0.985	0.985	0.983	0.965	0.977	0.975	0.984	0.984
	LIN	0.982	0.986	0.959	0.986	0.978	0.976	0.982	0.982
	PSM	0.990	0.988	0.974	0.977	0.976	0.973	0.986	0.986
4	MIX	0.939	0.930	0.924	0.913	0.935	0.925	0.956	0.956
	LIN	0.953	0.966	0.953	0.956	0.943	0.933	0.950	0.950
	PSM	0.945	0.956	0.954	0.943	0.938	0.924	0.954	0.954

MIX1 is a mixture model.

MIX2 is a mixture model with Prentice Criterion.

LIN1 is a random effects linear model.

LIN2 is a random effects linear model with a logit transformation.

LIN3 is a simple linear model with slope and intercept.

LIN4 is a simple linear model with no intercept. PSM is a principal stratification model.

Table 3

Standard error multiplier

Sample size	Data Set	MIX1	MIX2	LIN1	LIN2	LIN3	LIN4	PSM
original	1	1.24	1.53	1.39	1.46	1.38	1.36	1.21
	2	1.36	1.74	1.81	2.04	1.40	1.33	1.34
	3	1.51	1.58	2.06	1.89	1.23	1.25	1.50
	4	2.27	2.61	3.09	2.40	2.25	2.17	2.29
20% increase	1	1.17	1.46	1.33	1.40	1.31	1.29	1.13
	2	1.29	1.68	1.75	2.01	1.34	1.26	1.27
	3	1.45	1.52	2.02	1.85	1.21	1.22	1.44
	4	2.22	2.57	3.06	2.36	2.23	2.14	2.25

MIX1 is a mixture model.

MIX2 is a mixture model with Prentice Criterion.

LIN1 is a random effects linear model.

LIN2 is a random effects linear model with a logit transformation.

LIN3 is a simple linear model with slope and intercept.

LIN4 is a simple linear model with no intercept. PSM is a principal stratification model.