Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1046–1054. doi:10.1111/biom.12513.

Comparing Biomarkers as Trial Level General Surrogates

Erin E Gabriel^{1,*}, Michael J Daniels^{2,3,**}, and M Elizabeth Halloran^{4,5,6,***}

¹Biostatistics Research Branch, Division of Clinical Research, NIAID/NIH, Bethesda, MD USA

²Department of Statistics and Data Sciences, The University of Texas at Austin USA

³Department of Integrative Biology, The University of Texas at Austin USA

⁴Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA

⁵Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

⁶Center for Inference and Dynamics of Infectious Diseases, Seattle, WA, USA

Summary

An intermediate response measure that accurately predicts efficacy in a new setting can reduce trial cost and time to product licensure. In this paper, we define a *trial level general surrogate*, which is an intermediate response that can be used to accurately predict efficacy in a new setting. Methods for evaluating general surrogates have been developed previously. Many methods in the literature use trial level intermediate responses for prediction. However, all existing methods focus on surrogate evaluation and prediction in new settings, rather than comparison of candidate general surrogates, and few formalize the use of cross validation to quantify the expected prediction error. Our proposed method uses Bayesian non-parametric modeling and cross-validation to estimate the absolute prediction error for use in evaluating and comparing candidate trial level general surrogates. Simulations show that our method performs well across a variety of scenarios. We use our method to evaluate and to compare candidate trial level general surrogates in several multi-national trials of a pentavalent rotavirus vaccine. We identify at least one immune measure that has potential value as a trial level general surrogate and use it to predict efficacy in a new trial where the clinical outcome was not measured.

Keywords

Bayesian non-parametrics; Cross-Validation; Meta-analysis; Surrogate markers; Vaccines

1. Introduction

Of the many definitions of a surrogate the most common is that it is an intermediate response measurement that accurately predicts the clinical endpoint of interest. This definition lacks some necessary details about what ‘accurately predicts’ means and in what

*erin.gabriel@nih.gov

**mjdaniels@austin.utexas.edu

***betz@fhcrc.org

setting this prediction is done, but it is a useful starting point. Gilbert et al. (2008) refined the concept of a surrogate in the vaccine trial setting, proposing three levels: correlates, specific surrogates of protection, and general surrogates of protection. A correlate is a biomarker that is associated with the outcome in the vaccine arm or over both arms, and a specific surrogate of protection is a biomarker that is associated with efficacy. The definition of a general surrogate of protection is most closely related to the common definition of a surrogate: a biomarker on which the treatment effect can be used to accurately predict the treatment effect on the clinical outcome in a new setting. These levels are informative for general clinical trials as well. To make clear that we are considering the more general clinical trial setting, we will call biomarkers that are predictive of the clinical treatment effect in a new setting general surrogates (GS). Numerous papers have investigated the evaluation of biomarkers as GS, including Daniels and Hughes (1997); Gail et al. (2000); Burzykowski et al. (2001); Dai and Hughes (2012).

The definition of a GS in Gilbert et al. (2008) does not define in what new trial settings the GS will be useful or if the surrogate is at the individual or trial level. An individual level surrogate is a biomarker for which the individual level treatment effect on the biomarker is predictive of the individual level treatment effect on the clinical outcome. The effect of treatment on a trial level surrogate can be used to infer what the trial results would have been had the clinical outcome been measured. A GS can be predictive at the individual level or the trial level, or both. As outlined by Korn et al. (2005), these two types of surrogacy do not imply each other and can be unrelated. All existing methods for evaluating general surrogates, to our knowledge, consider only surrogate evaluation rather than comparison of candidate GS, and most are based on within sample prediction intervals. Only Baker (2006) attempts to formalize the exogenous quantification of the expected error when predicting in a new setting, using a summary of prediction error similar to our suggested method for a binary surrogate. As well, little attention is given to what information is available to support the generalizability of the association of the surrogate and the clinical endpoint to a new setting. In addition, most existing methods are specific for the type of data collected and require strict modeling assumptions. Dai and Hughes (2012) propose a general method for evaluating a trial level GS, but they require individual level data be available in all trials and suggest a within sample evaluation of prediction error based on a linear relationship between the treatment effects.

In this paper, we provide a precise definition of a trial level GS. We propose a general and flexible evaluation and prediction method that differs from existing meta-analytic evaluation methods in several ways. Our suggested estimation method can be used on individual level or trial level data, similar to Daniels and Hughes (1997). It allows for a flexible association between the treatment effect on the trial level GS and the treatment effect on the clinical outcome and flexible distributions for both true treatment effects over the trials. Our proposed evaluation method uses a Bayesian cross-validation approach to quantify the prediction error when estimating the treatment effect in a new setting. We propose the use of absolute prediction error as an error summary for evaluation and comparison of candidate trial level general surrogates. The absolute prediction error is easily interpretable because it is on the scale of the treatment effect on the clinical outcome. We build on the work of Daniels and Hughes (1997), taking a Bayesian approach to trial level GS evaluation while

allowing for the within trial estimation approach of Dai and Hughes (2012) when individual level data are available.

We use our method to evaluate and compare trial level GS for the pentavalent rotavirus vaccine RotaTeq™(RV5) (Merck & Co. Inc., Kenilworth, New Jersey). RV5 has been shown to be efficacious against rotavirus gastroenteritis in many settings, but in developing countries where the disease burden is highest, the efficacy is lower than in developed nations. Although some evidence of surrogacy has been found in previous work (Goveia, 2014), a clear measure of post-randomization immune response that explains the efficacy differences among the trials has yet to be identified. We show how our approach can be used to identify trial level GS in the setting of the RV5 trials.

In Section 2 we outline our proposed methods for estimation, prediction, evaluation and comparison. In Section 3 we explore the operating characteristics of our methods via several simulations studies, one of which closely follows the setting of the RV5 trials. In Section 4 we use our method to investigate trial level GS of rotavirus gastroenteritis of any severity in the RV5 trials, and in Section 5 we summarize our findings and suggest possible extensions.

2. Methods

2.1 Definitions and Notation

We refine the definition of a GS from Gilbert et al. (2008) to indicate that the prediction is at the trial level and characterize the new setting where such a GS might be used. If the trial level treatment effect on a biomarker can be used to predict (with low prediction error) a trial level clinical treatment effect and this predictive association is generalizable to a new setting, then the biomarker is a *trial level general surrogate* (TLGS) for that clinical outcome in that new setting.

This definition clearly states that the surrogate is at the trial level and emphasizes that the TLGS should only be used in a setting to which it is generalizable. When using a TLGS for prediction in a new setting, the justification for the evaluation trials as representative of the new setting should always be given. Justification for generalizability may include similar demographic, geographic, and clinical characteristics between the new trial and the evaluation trials.

Before describing our proposed methods for evaluation and comparison of biomarkers as TLGS, we define some notation. For subject $i \in \{1, \dots, n_j\}$ in trial $j \in \{1, \dots, J\}$, let $Z_{i,j} = \{0, 1\}$ be the treatment indicator, $Y_{i,j}$ the clinical endpoint, and $A_{i,j,k}$ the k th biomarker measure, $k \in \{1, \dots, K\}$. At the trial level, let $T_{c,j}$ be the true treatment effect on the clinical outcome in trial j and let $T_{s,j,k}$ be the true treatment effect on the k th biomarker measurement in trial j .

2.2 Within Trial Models

2.2.1 Models for Subject Level Data—When subject level data are available in one or more trials, we specify the following linear models for the (transformed) mean, median, or hazard for the candidate TLGS and the clinical outcome,

$$h_{s,k}(F_A[A_{i,j,k}|Z_{i,j}]) = \zeta_{0j,k} + T_{s,j,k}Z_{i,j}, \quad (1)$$

$$h_c(F_Y[Y_{i,j}|Z_{i,j}]) = \gamma_{0j} + T_{c,j}Z_{i,j}, \quad (2)$$

where $\zeta_{0j,k}$ and γ_{0j} denote the j th trial specific intercepts for the k th biomarker measurement and the clinical outcome, respectively. Models (1) and (2) are specified in terms of the distribution function, $F(\cdot)$, of A and Y and can be any set of models from which asymptotically normal coefficient estimates are obtained, including generalized linear models and survival models. We fit the models separately for each trial and response (outcome or biomarker). For this reason, the subjects in each trial used to fit (1) and (2), do not need to be the same provided the (implicit) missingness is random and/or by design. We denote the estimates of the treatment effects as $\hat{T}_{s,j,k}$ and $\hat{T}_{c,j}$ with standard errors, $\hat{\sigma}_{s,j,k}$ and $\hat{\sigma}_{c,j}$, $j \in \{1, \dots, J\}$.

Let $\mathbf{O}_{\mathbf{J},k}$ be the vector of estimated treatment effects and estimated standard errors over all J trials for the k th candidate TLGS, $\mathbf{O}_{\mathbf{J},k} = (\hat{T}_{s,1,k} \dots \hat{T}_{s,J,k}, \hat{T}_{c,1} \dots \hat{T}_{c,J}, \hat{\sigma}_{s,1,k} \dots \hat{\sigma}_{s,J,k}, \hat{\sigma}_{c,1} \dots \hat{\sigma}_{c,J})$; define $\mathbf{O}_{\mathbf{J}}$ to be the vector of all estimated treatment effects and standard errors over all K candidate GS and clinical outcome data in a given set of J trials. Also define $\mathbf{T}_c = (T_{c,1} \dots T_{c,J})$ and $\mathbf{T}_{s,k} = (T_{s,1,k} \dots T_{s,J,k})$.

Dai and Hughes (2012) use an estimating equation approach to estimate the joint sampling distribution of the estimated treatment effects from (1) and (2). Using their approach the correlation between the estimated treatment effects can be estimated. However, an estimate of this covariance is not required to evaluate the candidate TLGS (Daniels and Hughes, 1997), and missingness by design can make this covariance estimation complex.

2.2.2 Approximate Likelihood—We assume that the estimated treatment effects are consistent and asymptotically normal and can be well approximated by

$$\hat{T}_{s,j,k} \approx N(T_{s,j,k}, \hat{\sigma}_{s,j,k}^2) \text{ and } \hat{T}_{c,j} \approx N(T_{c,j}, \hat{\sigma}_{c,j}^2).$$

Note that the estimated treatment effects can be obtained by fitting models (1) and (2) or from the literature. We denote these models for the estimated treatment effects given the true treatment effects by $f_{s,j,k}(\hat{T}_{s,j,k}|T_{s,j,k}, \hat{\sigma}_{s,j,k})$ and $f_{c,j}(\hat{T}_{c,j}|T_{c,j}, \hat{\sigma}_{c,j})$. The joint distribution of the estimated treatment effects is given by

$$f_{j,k}(\hat{T}_{s,j,k}, \hat{T}_{c,j}|T_{s,j,k}, T_{c,j}, \hat{\sigma}_{s,j,k}, \hat{\sigma}_{c,j}) = f_{s,j,k}(\hat{T}_{s,j,k}|T_{s,j,k}, \hat{\sigma}_{s,j,k}) \times f_{c,j}(\hat{T}_{c,j}|T_{c,j}, \hat{\sigma}_{c,j}), \quad (3)$$

where we assume independence conditional on the true treatment effects. Based on these (approximate) trial specific models, the likelihood over all the trials for the k th TLGS is

given by $\prod_{j=1}^J f_{j,k}(\hat{T}_{s,j,k}, \hat{T}_{c,j} | \hat{\sigma}_{s,j,k}, \hat{\sigma}_{c,j}, T_{c,j}, T_{s,j,k})$, as the estimates between trials are assumed to be independent. In what follows, the estimated standard errors $\hat{\sigma}_{s,j,k}$ and $\hat{\sigma}_{c,j}$ will be treated as fixed.

When the individual level data is available, we can estimate the covariance of $(\hat{T}_{s,j,k}, \hat{T}_{c,j})$, and replace the two independent normals with a bivariate normal. One could also use the individual data directly by using models (1) and (2) to form the likelihood. However, using the individual level data is unlikely to substantially improve evaluation of GS at the trial level as was suggested in Korn et al. (2005). Therefore, we focus on a two step approach: first, estimate the treatment effects within each of a set of trials; and second, construct an approximate likelihood based on those estimates.

2.3 Between Trial Model and Priors

We specify a non-parametric prior for the distribution of the true treatment effect on the candidate GS, $T_{s,j,k}$. In particular we propose a Dirichlet process mixture of normals (MacEachern and Müller, 1998),

$$\begin{aligned} T_{s,j,k} &\sim N(\mu_{s,j,k}, \tau_{s,j,k}) \\ (\mu_{s,j,k}, \tau_{s,j,k}) &\sim G \\ G &\sim DP(\omega, G_0). \end{aligned} \quad (4)$$

For ω , we specify a $U(1, J)$ prior. We use this prior both to bound ω away from zero and to allow for clustering in the true trial level biomarker effects, $T_{s,j,k}$. We specify the base measure G_0 to be the product of a normal and a Gamma distribution where each $\mu_{s,j,k}$ follows a normal distribution with mean $\varphi_{s,k}$ and variance $\sigma_{\mu_{s,k}}^2$ and each $1/\tau_{s,j,k}$ follows a Gamma distribution with parameters $(\psi_{s,k}, \xi_{s,k})$. Denote the vectors of trial level parameters as $\boldsymbol{\mu}_{sk} = (\mu_{s,1,k}, \dots, \mu_{s,J,k})$ and $\boldsymbol{\tau}_{sk} = (\tau_{s,1,k}, \dots, \tau_{s,J,k})$. We use data-dependent hyper-priors for the parameters in the base measure as recommended in Taddy (2008).

We now specify a flexible model for the true trial level clinical treatment effect given the true trial level treatment effect on the TLGS, $[T_{c,j} | T_{s,j,k}]$. Specifically, $T_{c,j} = m(T_{s,j,k}, \beta_k, b_k) + \varepsilon_{j,k}$, where $\varepsilon_{j,k}$ are independent $N(0, \sigma_\varepsilon^2)$. The function $m(T_{s,j,k}, \beta_k, b_k)$ is defined as:

$$m(T_{s,j,k}, \beta_k, b_k) = \beta_{0,k} + \beta_{s,k} T_{s,j,k} + \sum_{q=1}^Q b_{q,k} |T_{s,j,k} - r_q|. \quad (5)$$

There are Q fixed knots/changepoints, r_q , such that $r_1 < r_q < \dots < r_Q$. The coefficients, $b_{q,k}$, associated with each knot are penalized/shrunk using $N(0, \kappa_{b_k}^2)$, with $1/\kappa_{b_k}^2$ given an Gamma $(1, 3)$ prior as recommended in Crainiceanu et al. (2005). The parameters $\beta_{0,k}$ and $\beta_{s,k}$ are

given independent $N(0, 106)$ priors. We use a diffuse Gamma prior for $1/\sigma_\varepsilon^2$. The full set of hyper-parameters will be denoted as

$$\nu = \{\tau_{sk}^2, \mu_{sk}, \varphi_{s,k}, \sigma_{\mu_{s,k}}^2, \psi_{s,k}, \xi_{s,k}, \mathbf{b}_k, \kappa_{b_k}^2, \sigma_\varepsilon^2, \beta_{0,k}, \beta_{s,k}, \omega\}.$$

To compare candidate TLGS, more than one candidate will need to be available in a given set of completed trials. When K candidates are available, we use an additive model for $[T_{c,j} | T_{s,j,1}, \dots, T_{s,j,K}]$, $T_{c,j} = \sum_k m_k(T_{s,j,k}, \beta_k, b_k) + \varepsilon_j$ where m has the form in (5) and ε_j are independent $N(0, \sigma_{\varepsilon_j}^2)$ with diffuse Gamma priors for $1/\sigma_{\varepsilon_j}^2$. The model containing the information from all K candidates is called the *full model*.

For the evaluation procedure presented in Section 2.4, we also need the null distribution of $\mathbf{T}_c$, i.e., the distribution of $\mathbf{T}_c$ independent of $\mathbf{T}_s$. We do not use the marginal $\mathbf{T}_c$ derived from the above models to avoid any potential model misspecification in the model for $\mathbf{T}_c | \mathbf{T}_s$. Instead, we specify the (null) prior for the true clinical treatment effects $\mathbf{T}_c$ ignoring all candidate TLGS information using the same specification and priors as in (4).

2.4 Evaluation and Comparison

Our objective is to compare and evaluate candidate TLGS based on their predictive power and then use the best TLGS to predict the true clinical treatment effect in a new trial, denoted by $J+1$, where the estimated treatment effect on the clinical outcome, $\hat{T}_{c,J+1}$, is not observed. For prediction we use the posterior distribution of $T_{c,J+1}$, the true treatment effect on the clinical outcome in the new trial $J+1$, given the observed data, which includes $\hat{T}_{s,J+1,k}$ and $\hat{\sigma}_{s,J+1,k}$ in addition to $\mathbf{O}_{J,k}$, i.e., $T_{c,J+1} | \mathbf{O}_{J,k}, \hat{T}_{s,J+1,k}, \hat{\sigma}_{s,J+1,k}$. To evaluate predictive accuracy and compare candidate TLGS, we will use the expected absolute prediction error,

$$D_{J+1,k} = E |T_{c,J+1} - \hat{T}_{c,J+1,k}^*|. \quad (6)$$

The leave-one-out prediction of $T_{c,j}$ based on the k th candidate TLGS, $\hat{T}_{c,j,k}^*$, is computed from the posterior obtained by excluding $\hat{T}_{c,j}$ and $\hat{\sigma}_{c,j}$ as well as all information about all other candidate TLGS, from the data, $\hat{T}_{c,j,k}^* = E[T_{c,j} | \mathbf{O}_{J(-j),k}, \hat{T}_{s,j,k}, \hat{\sigma}_{s,j,k}]$, where $\mathbf{O}_{J(-j),k}$ is the information about the clinical outcome and the k th candidate GS in all $\{1, \dots, J\}$ trials less trial j .

The summary $D_{J+1,k}$ is the expected absolute prediction error for a new trial $J+1$. If we observed the true $T_{c,j}$, we could estimate the absolute prediction error for each observed trial j by comparing a leave-one-out estimate, $\hat{T}_{c,j,k}^*$, to the true clinical treatment effect $T_{c,j}$, which we will call $\hat{d}_{j,k}$, and then average over the J trials, obtaining:

$$\hat{D}_{J+1,k} = (1/J) \sum_j \hat{d}_{j,k} = (1/J) \sum_j |T_{c,j} - \hat{T}_{c,j,k}^*|. \quad (7)$$

Similar cross-validation estimates were found to have good finite sample properties for independent and identically distributed (iid) data in Tian et al. (2007).

Since we will never observe the true $T_{c,j}$, even in the evaluation trials, we cannot use (7) to estimate (6). Instead, we will use an approximation to the distribution of (7) to assess candidate TLGS. In particular, we use

$$\tilde{d}_{j,k}|\mathbf{O}_J \sim |T_{c,j} - \hat{T}_{c,j,k}^*|, \quad (8)$$

the posterior distribution of the approximate absolute prediction error for trial j using all the data, $\mathbf{O}_J$. By taking the expected value of (8), i.e., with respect to the posterior distribution, $T_{c,j}|\mathbf{O}_J$, we obtain an approximation of $\hat{d}_{j,k}$ in (7). For evaluation of a TLGS in an arbitrary new setting, we use the mixture distribution, $\tilde{D}_{J+1,k} \sim (1/J)\sum_j \tilde{d}_{j,k}$. The expectation of $\tilde{D}_{J+1,k}$ is an approximation of (7) and is a single value comparison for candidate TLGS in a new setting.

The distribution of $\tilde{D}_{J+1,k}$ is of interest for comparing candidate TLGS. As $\tilde{D}_{J+1,k}$ is on the scale of the true clinical effect, a first step in TLGS evaluation can be a comparison of $\tilde{D}_{J+1,k}$ to the smallest clinically relevant effect, as was suggested in Baker (2006). However, to evaluate the absolute quality of a candidate TLGS we will compare $\tilde{D}_{J+1,k}$ to the absolute prediction error in the absence of $\hat{T}_{s,j,k}$; we will denote this as $\tilde{D}_{J+1,0}$. We use the null model prior for the distribution of the true $T_{c,j}$ introduced at the end of Section 2.3. Similar to the development for $\tilde{d}_{j,k}$, we use the following approximate distribution of the absolute prediction error for the null model,

$$\tilde{d}_{j,0}|\mathbf{O}_J \sim |T_{c,j} - E[T_{c,j}|\mathbf{O}_{J(-j,-s)}]|, \quad (9)$$

where $\mathbf{O}_{J(-j,-s)}$ denotes the observed clinical outcome information alone, less that in trial j , i.e., $\mathbf{O}_{J(-j,-s)} = (\hat{T}_{c,q}, \hat{\sigma}_{c,q}) : q \in \{1, \dots, j-1, j+1, \dots, J\}$. The distribution $\tilde{D}_{J+1,0} \sim (1/J)\sum_j \tilde{d}_{j,0}$ is a good baseline for candidate TLGS evaluation because it quantifies the amount of error that occurs when no potential TLGS are used in the model.

To evaluate and to compare candidate GS based on $\tilde{D}_{J+1,k}$ and $\tilde{D}_{J+1,0}$, a joint graphical representation of the density estimates can be useful. Densities of $\tilde{D}_{J+1,k}$ that have more mass at 0 with shorter tails are evidence of superior candidates, as this implies greater probability of lower prediction error. Figure 1 is an illustration of how the density plots can be used to compare and to evaluate candidate TLGS. Both candidate TLGS in Figure 1 are superior to the null model, while the density of $\tilde{D}_{J+1,l}$ in comparison to that of $\tilde{D}_{J+1,k}$ suggests that candidate l is superior to candidate k . In this example, candidate l is an ideal TLGS, with a true $\tilde{D}_{J+1,l} = 0$ as we use the true treatment effects $T_{c,j}$ as $T_{s,j,l}$ for each trial. However, the k th candidate is still a very useful TLGS.

The probability $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,k} | \mathbf{O}_J)$, can be used to formally quantify the strength of evidence that a given candidate TLGS has any value. Small probabilities suggest evidence that the k th candidate has value as a TLGS to aid in the prediction of the clinical treatment effect in a new trial setting. One can also rank candidates using the set of point and interval estimates from the set of $\{\tilde{D}_{J+1,1}, \dots, \tilde{D}_{J+1,K}\}$ distributions. The probability $P(\tilde{D}_{J+1,k} < \tilde{D}_{J+1,l} | \mathbf{O}_J)$, can be used to quantify the strength of evidence for the superiority of a given candidate over another within the same set of evaluation trials. Here, small probabilities suggest that there is evidence of the superiority of the l th TLGS candidate over the k th candidate, while large probabilities provide evidence of the opposite. Posterior probabilities near 0.5 in either $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,k} | \mathbf{O}_J)$ or $P(\tilde{D}_{J+1,k} < \tilde{D}_{J+1,l} | \mathbf{O}_J)$ are evidence of a useless surrogate or two TLGS with similar usefulness, respectively.

2.5 Posterior computation

Computation of all quantities of interest requires four steps, which we outline below.

1. Sample from the posterior of the full model using all observed data on all K candidate TLGS to obtain the marginal posterior of each $T_{c,j}$. This step provides the best estimate of the true value of the treatment effect on the clinical outcome.
2. Removing the observed clinical data information for trial j and all but one of the K candidate TLGS, we sample from the corresponding posterior to obtain the leave-the- j th-trial-out-estimate, $\hat{T}_{c,j,k}^*$. We repeat for all J evaluation trials and all K candidate TLGS.
3. Using only the clinical information from the other trials besides trial j and removing all candidate surrogate information, we sample from the corresponding posterior to obtain an estimate of the true treatment effect on the clinical outcome, $E[T_{c,j} | \mathbf{O}_{J(-j-s)}]$. We repeat this for all J evaluation trials.
4. We use the posteriors of $\mathbf{T}_c$ from steps 1 and 2 to obtain the distributions of $\tilde{D}_{J+1,k}$ for each candidate TLGS and the posterior distributions of $\mathbf{T}_c$ from steps 1 and 3 to obtain the distribution of $\tilde{D}_{J+1,0}$.

We use the posterior from step 2 to estimate the true clinical treatment effect in a trial setting where only the candidate surrogate is measured using a given TLGS.

3. Simulations

We investigate the performance of our method in several scenarios, including for different relationships between the true treatment effect on the clinical outcome, $T_{c,j}$, and the true treatment effect on the surrogate, $T_{s,j,k}$ and for different distributions governing the true $T_{s,j,k}$, including a single normal or a mixture of normals. For each of these scenarios we investigate the impact of the magnitude of the errors with which the treatment effects were estimated within the trials, $\hat{\sigma}_{s,j,k}$ and $\hat{\sigma}_{c,j}$. For the scenarios where $T_{s,j,k}$ are generated from a normal distribution, we fix the mean and standard deviation (SD) of each $T_{s,j}$ to be 2 and 1, respectively. In the mixture of normals settings, we generate $T_{s,j,k}$ from an equally weighted

mixture of 3 normals with means $(-1, 0, 2)$ and SDs of 1. We fix the SD of the generated $T_{c,j}$, conditional on $T_{j,1,k}$, to be 1.

For each scenario we simulate 200 datasets of 20 trials, unless otherwise specified. We investigate four scenarios with fewer than 20 trials, two with five trials and one each with ten and twelve trials. The exact specifications of the simulation scenarios are given in the supplementary materials; the R and JAGS code used to generate the data and fit the models is also available in the supplementary materials (R Core Team, 2015; Plummer, 2003, 2015). The last scenario in Tables 1 and 2 is based on the rotavirus vaccine trial application, with 12 trials and using the set of $\hat{T}_s$ and $\hat{\sigma}_s$ from one of the candidate TLGS as well as assuming a strong linear association between the true treatment effects. For our method outlined in Section 2, we use a piecewise linear model as given in (5) with knots at the 33rd and 66th percentiles of the $\hat{T}_{s,j,k}$ for the prior on $T_{c,j}|T_{s,j,k}$. We compare our method's predictive accuracy, as described in Section 2.4, with 1) a linear regression on the *estimated* treatment effects (referred to as 'linear model' in what follows) and 2) the method of Daniels and Hughes (1997) (referred to as DH in what follows).

Table 1 reports the mean and SD of $\hat{D}_{J+1,k} = (1/J) \sum_j [T_{c,j}^{true} - E[T_{c,j} | \mathbf{O}_{J(-j)}, \hat{T}_{s,j,k}, \hat{\sigma}_{s,j,k}]]$. $T_{c,j}^{true}$ is the true clinical treatment effect for trial j , where J is the number of trials. Table 1 also displays the mean and bias of the posterior mean of $\tilde{D}_{J+1,k}$, given by (8), in comparison to $\hat{D}_{J+1,k}$, given by (7). Both the DH method and our proposed method estimate $\hat{D}_{J+1,k}$ more accurately than the linear model in all scenarios. The DH method and our proposed method have similar bias. The $\tilde{D}_{J+1,k}$ are always lower for our proposed method in non-linear scenarios and similar in linear scenarios for our suggested method over the DH method. However, in many cases the linear model method has the lowest $\hat{D}_{J+1,k}$, suggesting it has good out-of-sample prediction. Table S1 of the supplementary materials reports the root mean squared error and the bias based on the true clinical treatment effects, which cannot be observed in practice, and the within sample and leave-one-out estimates based on the three methods. These errors show that although all the methods can give good out-of-sample predictions when the model is correct, only the two more flexible methods have adequate in-sample estimation to unbiasedly approximate $\hat{D}_{J+1,k}$ by $\tilde{D}_{J+1,k}$. Table S2 of the supplementary materials shows the Monte Carlo error for the differences in $\hat{D}_{J+1,k}$ and $\tilde{D}_{J+1,k}$ between the methods; these are small relative to the differences between the methods.

Table 2 reports the average $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,k} | \mathbf{O}_J)$ over the 200 simulated datasets for each scenario and method. For both the DH method and our proposed method the average probabilities are near 0.5 in scenarios where there is no useful candidate surrogate, i.e. the first row of Table 2. In addition, the lower probabilities in all other scenarios suggest that this probability can help quantify the value of a candidate as a TLGS. The low probabilities for all scenarios for the linear model are caused by the biased estimation of $D_{J+1,k}$, which makes $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,k} | \mathbf{O}_J)$ under this method non-informative for evaluating TLGS candidates. Table 2 also presents the average probability for comparing two candidates, $P(\tilde{D}_{J+1,k} < \tilde{D}_{J+1,l} | \mathbf{O}_J)$, where $\tilde{D}_{J+1,l}$ is the estimate based on a TLGS with a true $\hat{D}_{J+1,l} = 0$, from using $T_{c,j}^{true}$ as the estimated treatment effect on the candidate GS. The estimates of

$P(\tilde{D}_{J+1,l} < \tilde{D}_{J+1,k} | \mathbf{O}_J)$ being larger than 0.5 in almost every case suggests that this probability can be used to discern between candidate TLGS of differing quality.

Table S3 in the supplementary materials summarizes the simulations using subject level data (as opposed to trial level summaries directly) to compare the two different ways of formulating the within-trial likelihoods. We find that the methods are basically comparable, with the working independence version of the likelihood having lower bias in the estimation of $\hat{D}_{J+1,k}$ by $\tilde{D}_{J+1,k}$, although the bivariate normal likelihood method tends to have slightly smaller $\tilde{D}_{J+1,k}$ on average.

4. Application: Pentavalent Rotavirus Vaccine

More than 450,000 children under five years died from complications of rotavirus infection each year (Tate et al., 2012) prior to vaccine availability. The pentavalent rotavirus vaccine RotaTeq™(RV5) developed by Merck & Co. Inc. has been licensed for use in over 120 countries. The Rotavirus Efficacy and Safety Trial (REST) against severe rotavirus gastroenteritis (Vesikari et al., 2006) was conducted in 11 countries. In the sub-study of REST in Finland and the United States, estimated efficacy was as high as 98%. However, in other regions in Africa and Asia, lower efficacy has been observed (Armah et al., 2010; Zaman et al., 2010). Lower efficacy may be related to difference in participants' immune system function. If this is the case, such measurements could be used to better predict efficacy in future settings.

We investigate several candidate TLGS using data from four phase II and III studies of RV5 conducted in seven countries: Finland, the United States, Vietnam, Mali, Bangladesh, Ghana and Kenya (Vesikari et al., 2006; Heaton et al., 2005; Armah et al., 2010, 2012; Zaman et al., 2010; Shin et al., 2012). Participants from Finland and the United States were included in more than one study. Some of the trials involved more than one country. For the purpose of this analysis, we assume participants from different countries in the same trial to be independent trials as it is unlikely that outcomes will be correlated between countries even within the same trial setting. Table 3 shows the 12 trials (based on study and country) used as our evaluation trials, and a 13th trial in Taiwan where only the potential TLGSs are measured. The data were provided by Merck & Co. Inc., Kenilworth, New Jersey, through data sharing agreements with the Fred Hutchinson Cancer Research Center and the National Institute of Allergy and Infectious Diseases.

For this application we consider rotavirus gastroenteritis of any severity as the outcome of interest. Dose, endemic burden of disease, age range, ethnicity, and region were available for each trial. Several candidate TLGS were measured in a subset of individuals. These included serum anti-rotavirus IgA B-cell responses (IgA), as well as serum neutralization antibody (SNA) to the human rotavirus serotypes G1, G2, G3, G4 and P1A. Given that the different trials collected all the markers on only a random sample of those with the clinical outcome, we did not use the method of Dai and Hughes (2012), but instead present results using the working-independence model (3) based on estimates obtained from fitting independent generalized linear models in each trial for the available immunogenicity and clinical outcome data.

Figure 2 displays the point estimates and 95% credible intervals (CI) for the vaccine effect on the clinical outcome (rate of gastroenteritis of any severity), $T_{c,j}$, and the vaccine effect on two selected immune markers, SNA G1 and serum anti-rotavirus IgA, $T_{s,j,k}$. Figure S1 in the supplementary materials describes this same relationship for all the other biomarkers collected. The curve overlaid in Figure 2, and Figure S1 of the supplement, is the posterior mean from the regression model given in (5). As all biomarkers showed at least some level of surrogate value, it is likely a combination of the biomarkers will have the most value as a TLGS.

The $T_{c,j}$ estimates and standard errors are obtained by fitting Poisson regression models, including an offset for the log of follow-up time. This is the same model used in the clinical papers on these trials (Heaton et al., 2005; Vesikari et al., 2006; Armah et al., 2010, 2012; Zaman et al., 2010; Shin et al., 2012). The treatment effect on the clinical outcome in each trial is the difference in log rates of rotavirus gastroenteritis of any severity between the vaccinated and unvaccinated children. A linear model was used to estimate the vaccine effect on the two potential TLGS with the effect being the difference in the log titer level between the vaccinated and unvaccinated participants.

For ease of notation let $k = \{G1, IGA\}$. The mean of the predictive error distribution, $D_{J+1,G1}$ is 0.29 with a 95th percentile of 0.8; for $D_{J+1,IGA}$, 0.36, with a 95th percentile of 1.03. We find evidence that both serum anti-rotavirus IgA and SNA G1 have some value as TLGS for rotavirus gastroenteritis of any severity. This can be seen in Figure 3 and from the probabilities $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,G1} | \mathbf{O}_J) = 0.19$ and $P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,IGA} | \mathbf{O}_J) = 0.22$ being less than 0.5. The probability $P(\tilde{D}_{J+1,G1} > \tilde{D}_{J+1,IGA} | \mathbf{O}_J) = 0.46$ suggests the two candidates are similar but that there is some evidence that G1 is superior. These results can also be seen in Figure 3, as both $\tilde{D}_{J+1,IGA}$ and $\tilde{D}_{J+1,G1}$ have more mass closer to 0 and less spread than $\tilde{D}_{J+1,0}$.

Only selected sites had the clinical outcome measure collected in the REST clinical trial (Vesikari et al., 2006). Immune measurements alone were taken at the Taiwan site. We estimate the true clinical vaccine effect at the Taiwan site, $T_{c,Taiwan}$ ('Taiwan' corresponds to $J+1$) to be -1.51 using G1 as a TLGS with 95% CI $(-2.3, -0.90)$; this corresponds to a vaccine efficacy estimate of $1 - \exp(-1.51) = 78\%$, with 95% CI (59%, 90%) against rotavirus gastroenteritis of any severity. Given the 95th percentile of the distribution $\tilde{D}_{J+1,G1}$ and the the upper 95% CI limit of the credible interval for $T_{c,Taiwan}$, positive efficacy in Taiwan is likely $(-0.9+0.8) < 0$. Figure S2 of the supplementary materials illustrates this graphically, giving a maximum prediction interval based on $\tilde{D}_{J+1,G1}$ of $(-3.1, -0.1)$ for the clinical vaccine effect, which gives a vaccine efficacy interval of (9.5%, 95%). The subjects in the Taiwan study come from the same general region, ethnic group, age range and socioeconomic status as many of the subjects in the 12 evaluation trials. Therefore, we believe the predictive relationship between the vaccine effect on G1 and the vaccine effect on rotavirus gastroenteritis of any severity is likely to generalize.

5. Discussion

We provide a definition for a trial level general surrogate and outline a procedure for the evaluation and comparison of TLGS candidates using the summary absolute prediction error. We propose a non-parametric Bayesian model for the prediction of clinical treatment effects in a new setting. We demonstrate that this model has good predictive properties, particularly when the true relationship between the treatment effects is non-linear. Our proposed evaluation procedure can use separate models for the 'full-model' estimation and prediction. This allows for the comparison of predictive models as well as candidate TLGS. The approach is not restricted to use with the absolute prediction error. Other loss/error functions can be used by simply replacing the absolute prediction error in the evaluation stage.

In the application, we find evidence of a useful TLGS for rotavirus gastroenteritis of any severity in the RV5 vaccine trials. Previous studies have found that serum anti-rotavirus IgA has value as a GS (Goveia, 2014). We plan to investigate and compare TLGS for this setting more extensively in a clinical manuscript. Using our evaluated TLGS, we predict that positive clinical efficacy would have been observed in the Taiwan study, a setting to which we believe the TLGS based prediction is generalizable. As is demonstrated in the application and pointed out in Section 2.2.1, our proposed method allows the use all available outcome and immunogenicity data in each trial to estimate the treatment effect on the clinical outcome and the treatment effect on the candidate GS, even if this information is not for the same set of subjects.

Useful extensions to our proposed method would be the simultaneous evaluation of the individual level GS and TLGS, as discussed in Alonso et al. (2015) or the evaluation of surrogate consistency, as discussed and demonstrated for a specific meta-analytic setting in Elliott et al. (2015). The consideration of combinations of measures as surrogates is also of interest for future research.

6. Supplementary Materials

Web Appendices, Tables, Figures and The R and JAGS code referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Merck & Co. Inc. (Kenilworth, New Jersey) for sharing their RotaTeq™ clinical trial data with us. The authors received no funding from Merck & Co. Inc. for this analysis nor did Merck & Co. Inc. influence the analysis in any way. M.J. Daniels was partially supported by National Institutes of Health R01 CA183854. M.E. Halloran was partially supported by National Institutes of Health R37 AI032042. We would also like to acknowledge Peter Gilbert, Michael Sachs and James Dai for their helpful ideas, program ming expertise and discussions. Peter Gilbert is partially supported by NIAID NIH Award Number R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

References

- Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*. 2015; 71:15–24. [PubMed: 25274284]
- Armah GE, Breiman RF, Tapia MD, Dallas MJ, Neuzil KM, Binka FN, Sow SO, Ojwando J, Ciarlet M, Steele AD. Immunogenicity of the pentavalent rotavirus vaccine in African infants. *Vaccine*. 2012; 30(Supplement 1):A86–A93. [PubMed: 22520142]
- Armah GE, Sow S, Breiman RF, Dallas MJ, Tapia MD, Feikin DR, Binka FN, Steele AD, Laserson KF, Ansah NA, Levine MM, LK, Coia ML, Attah-Poku M, Ojwando J, Rivers SB, Victor JC, Nyambane G, Hodgson A, Schdel F, Ciarlet M, Neuzil KM. Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in Sub-Saharan Africa: a randomized, double-blind, placebo-controlled trial. *The Lancet*. 2010; 376:606–614.
- Baker SG. A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. *Biostatistics*. 2006; 7:58–70. [PubMed: 15972889]
- Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2001; 50:405–422.
- Crainiceanu C, Ruppert D, Wand M. Bayesian analysis for penalized spline regression using WinBugs. *Journal of Statistical Software*. 2005; 14:1–24.
- Dai JY, Hughes JP. A unified procedure for meta-analytic evaluation of surrogate end points in randomized clinical trials. *Biostatistics*. 2012; 13:609–624. [PubMed: 22394448]
- Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*. 1997; 16:1965–1982. [PubMed: 9304767]
- Elliott MR, Conlon AS, Li Y, Kaciroti N, Taylor JM. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics*. 2015; 16:400–412. [PubMed: 25236906]
- Gail M, Pfeiffer R, van Houwelingen H, Carroll R. On meta-analytic assessment of surrogate outcomes. *Biostatistics*. 2000; 1:231–246. [PubMed: 12933506]
- Gilbert P, Qin L, Self S. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*. 2008; 27:4758–4778. [PubMed: 17979212]
- Goveia, M. Update on pentavalent human-bovine rotavirus vaccine. Presentations from the 11th International Rotavirus Symposium; 2014.
- Heaton PM, Goveia MG, Miller JM, Offit P, Clark HF. Development of a pentavalent rotavirus vaccine against prevalent serotypes of rotavirus gastroenteritis. *Journal of Infectious Diseases*. 2005; 192:S17–S21. [PubMed: 16088800]
- Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine*. 2005; 24:163–182. [PubMed: 15515150]
- MacEachern SN, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*. 1998; 7:223–238.
- Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003. <http://mcmc-jags.sourceforge.net/>
- Plummer, M. R package version 3-15. 2015. rjags: Bayesian Graphical Models using MCMC.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2015. <https://www.R-project.org/>
- Shin S, Anh DD, Zaman K, Yunus M, Mai LTP, Thiem VD, Azim T, Victor JC, Dallas MJ, Steele AD, Neuzil KM, Ciarlet M. Immunogenicity of the pentavalent rotavirus vaccine among infants in two developing countries in Asia, Bangladesh and Vietnam. *Vaccine*. 2012; 30(Supplement 1):A106–A113. Rotavirus Vaccines for Children in Developing Countries. [PubMed: 22520119]
- Taddy, MA. PhD thesis. University of California; Santa Cruz: 2008. Bayesian Nonparametric Analysis of Conditional Distributions and Inference for Poisson Point Processes.
- Tate JE, Burton AH, Boschi-Pinto C, Steele AD, Duque J, Parashar UD. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal

rotavirus vaccination programs: a systematic review and meta-analysis. *The Lancet Infectious Diseases*. 2012; 12:136–141. [PubMed: 22030330]

Tian L, Cai T, Goetghebeur E, Wei LJ. Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*. 2007; 94:297–311.

Vesikari T, Matson DO, Dennehy P, Van Damme P, Santosham M, Rodriguez Z, Dallas MJ, Heyse JF, Goveia MG, Black SB, Shinefield HR, Christie CD, Ylitalo S, Itzler RF, Coia ML, Onorato MT, Adeyi BA, Marshall GS, Gothefors L, Campens D. Safety and efficacy of a pentavalent human-bovine (wc3) reassortant Rotavirus vaccine. *New England Journal of Medicine*. 2006; 354:23–33. [PubMed: 16394299]

Zaman K, Anh DD, Victor JC, Shin S, Yunus M, Dallas MJ, Podder G, Thiem VD, Mai LTP, Luby SP, Tho LH, Coia ML, Lewis K, Rivers SB, Sack DA, Schödel F, Steele AD, Neuzil KM, Ciarlet M. Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in Asia: a randomised, double-blind, placebo-controlled trial. *The Lancet*. 2010; 376:615–623.

Figure 1.

Density estimates for $\tilde{D}_{J+1,k}$, $\tilde{D}_{J+1,l}$ and $\tilde{D}_{J+1,0}$. Candidate TLGS k is a good, but not perfect TLGS. Candidate l is an ideal TLGS, with a true $\hat{D}_{J+1,l} = 0$. $\tilde{D}_{J+1,0}$ corresponds to the model containing no candidate TLGS. Large mass closer to 0, and less spread away from zero, for $\tilde{D}_{J+1,k}$ and $\tilde{D}_{J+1,l}$ suggests a smaller mean and lower variance of the approximate distribution of the absolute prediction error.

Figure 2.

Association of Estimated Vaccine Effects. Each sub-figure displays the relationship between the estimated treatment effect on the candidate trial level GS, $\hat{T}_{s,k,j}$ on the x-axis, and the estimated treatment effect on the clinical outcome, $\hat{T}_{c,j}$ on the y-axis, over the 12 trial settings. The two candidates of interest are G1 or IgA. The crossbars depict the 95% CI for the estimates. The curve overlaid is the posterior mean from the spline model given in (5). The units are difference in log rates for both the X and Y axis.

Figure 3.
Density estimates for $\tilde{D}_{J+1,IGA}$, $\tilde{D}_{J+1,G1}$ and $\tilde{D}_{J+1,0}$

Mean $\hat{D}_{j+1,k}$ with (SD) and Mean and Bias of $\tilde{D}_{j+1,k} \cdot \hat{D}_{j+1,k} = (1/J) \sum_j |T_{c,j}^{true} - E[T_{c,j} | O_{J(-)}, \hat{T}_{s,j}, \hat{\sigma}_{i,j}]|$ where $T_{c,j}^{true}$ is the true clinical treatment effect simulated for trial j. DH refers to the method used in Daniels and Hughes (1997). Each scenario has 200 replicates.

Table 1

Scenario	Mean $\hat{D}_{j+1,k}$ (SD over Replications)			Mean $\tilde{D}_{j+1,k}$ Bias		
	Linear Model	DH	New Method	Linear Model	DH	New Method
0, Linear no-effect	0.87 (0.16)	1.15 (0.15)	1.27 (0.18)	0.09, 0.656	1.15, 0.001	1.28, -0.002
1, Linear	0.10 (0.02)	0.48 (0.07)	0.16 (0.03)	0.01, 0.078	0.49, -0.004	0.15, 0.003
2, Linear large SE $\hat{T}_s$	0.67 (0.11)	1.07 (0.15)	1.05 (0.16)	0.07, 0.513	1.07, -0.001	1.05, -0.001
3, Linear large SE $\hat{T}_c$	0.30 (0.11)	0.73 (0.09)	0.54 (0.10)	0.09, 0.193	0.91, -0.132	0.59, -0.029
4, Linear large Beta	0.09 (0.01)	0.48 (0.08)	0.14 (0.01)	0.01, 0.066	0.49, -0.004	0.14, 0.001
5, Linear 10 trials	0.13 (0.04)	0.54 (0.12)	0.23 (0.06)	0.03, 0.088	0.55, -0.004	0.23, 0.003
6, Linear 5 trials	0.18 (0.07)	0.80 (0.25)	0.51 (0.23)	0.09, 0.083	0.81, -0.002	0.50, 0.005
7, Cubic-threshold	0.75 (0.25)	1.01 (0.29)	0.36 (0.19)	0.11, 0.562	1.01, -0.001	0.36, -0.001
8, Step	0.99 (0.12)	1.16 (0.09)	0.84 (0.18)	0.11, 0.802	1.16, 0.002	0.84, 0.001
9, Cubic	0.41 (0.20)	0.69 (0.23)	0.18 (0.09)	0.06, 0.311	0.69, -0.002	0.18, 0.000
10, Linear, non-iid T_s	0.15 (0.03)	0.39 (0.05)	0.22 (0.04)	0.02, 0.113	0.40, -0.004	0.22, 0.003
11, Step, non-iid T_s	0.99 (0.13)	1.13 (0.09)	0.82 (0.21)	0.11, 0.836	1.13, 0.003	0.82, 0.000
12, Step, non-iid T_s , 5 trials	1.47 (0.64)	1.94 (0.83)	1.82 (0.66)	0.76, 0.683	1.94, 0.006	1.81, 0.006
13, Example scenario, 12 trials	0.50 (0.09)	0.79 (0.14)	0.73 (0.16)	0.09, 0.365	0.81, -0.009	0.68, 0.045

Detection of predictive power based on $\tilde{D}_{J+1,k}$

Table 2

$\tilde{D}_{J+1,0}$ is the distribution of the prediction error for the null model containing no GS candidates. $\tilde{D}_{J+1,k}$ is the distribution of the prediction error for the candidate created in each scenario, and $\tilde{D}_{J+1,l}$ is the distribution of the prediction error for the TLGS with a true $\hat{D}_{J+1,l} = 0$, using $T_{c,j}^{true}$ as the candidate TLGS. DH refers to the method used in Daniels and Hughes (1997). Each scenario has 200 replicates.

Scenario	$P(\tilde{D}_{J+1,0} < \tilde{D}_{J+1,k} \mathbf{O}_J)$		$P(\tilde{D}_{J+1,l} < \tilde{D}_{J+1,k} \mathbf{O}_J)$	
	Linear Model	DH	New Method	DH
0, Linear no-effect	0.033	0.424	0.452	0.963
1, Linear	0.005	0.227	0.075	0.636
2, Linear large SE $\hat{T}_s$	0.029	0.429	0.414	0.947
3, Linear large SE $\hat{T}_c$	0.038	0.374	0.255	0.197
4, Linear large Beta	0.004	0.222	0.064	0.624
5, Linear 10 trials	0.011	0.217	0.090	0.674
6, Linear 5 trials	0.022	0.250	0.147	0.527
7, Cubic-threshold	0.012	0.166	0.064	0.961
8, Step	0.087	0.495	0.399	0.994
9, Cubic	0.036	0.450	0.127	0.895
10, Linear, non-iid T_s	0.006	0.177	0.097	0.710
11, Step, non-iid T_s	0.123	0.507	0.417	0.997
12, Step, non-iid T_s , 5 trials	0.199	0.427	0.371	0.906
13, Example scenario, 12 trials	0.037	0.317	0.270	0.576

Table 3

Data available for each trial used in the analysis in Section 5.

Protocol Number	Region/ Grouping	Country	Number of Subjects with		
			Clinical Outcome	IgA	G1
015	Asia	Bangladesh	1116	146	146
005 [‡]	Finland	Finland	1027	644	647
006	Finland	Finland	2324	358	1503
007	Finland	Finland	637	54	54
015	Africa	Ghana	1971	143	143
015	Africa	Kenya	1137	128	128
015	Africa	Mali	1667	137	137
006	Native lands	United States	583	207	207
006	US concomitant	United States	1239	106	104
006	US non-concomitant	United States	366	220	210
007	United States	United States	478	97	98
015	Asia	Vietnam	871	149	149
006	Asia	Taiwan	0	99	99

[‡] Only the placebo and low, middle, and high dose RV5 groups were included.