Evaluation Methods for T-Association of a Surrogate Endpoint

Jo-Ying Hung[1], Chih-Yuan Hsu[2], Pei-Fang Su[1], and Yu Shyr[2*]

[1]Department of Statistics, National Cheng Kung University, Tainan, 70101, Taiwan
[2,]Department of Biostatistics, Vanderbilt University Medical Center

August 28, 2025

Abstract

A surrogate endpoint is a biomarker or intermediate clinical outcome used instead of a direct clinical endpoint to predict the clinical benefit of a drug. It serves as a substitute for a primary endpoint, offering benefits when it can be measured earlier or more conveniently. Before using a surrogate endpoint for scientific conclusions, its qualification must be assessed. A valid surrogate endpoint must meet two associations: I-Association (the association between the surrogate and true endpoints, such as disease response and overall survival) and T-Association (the association between treatment effects on both endpoints, such as odds ratio and hazard ratio). I-association is commonly evaluated, but T-association is often overlooked. This study proposes methods to assess T-association since both treatment effects are estimated as two random variables. We assume that the treatment effects on the surrogate and true endpoints follow a bivariate normal distribution, where the key evaluation metric is the correlation coefficient, which quantifies the relationship between treatment effects on both endpoints. The model parameters, including the correlation coefficient of interest, are estimated using maximum likelihood, restricted maximum likelihood, and a Bayesian approach. Simulated data and real-world data are used to demonstrate these methods. In simulations, we evaluate bias, standard error, and coverage probability. This method will serve as the statistical foundation for future FDA accelerated approval drugs.

Keywords: Surrogate endpoint, Bivariate normal, maximum likelihood, restricted maximum likelihood, Bayesian

*Corresponding author. Email: yu.shyr@vumc.org

1 Introduction

The process of getting a new drug approved by the U.S. Food and Drug Administration (FDA) typically takes a long time from initial investigation to final authorization. To quantify how long the process takes, Brown et al. (2022) analyzed 268 drugs approved from 2010 to 2019 and found that the mean duration between the submission of an investigational new drug application and the approval was 8.7 years. This lengthy process can potentially lead to delay in clinical treatment, causing patients to miss timely access to effective therapies. For instance, Stewart et al. (2018) estimated that shortening the process by one year could save a median of 79,920 life-years per drug, indicating the substantial benefit if the drug development efficiency was improved.

There are several factors that can prolong the duration of clinical trials, such as patient recruitment difficulties, the number of events required for analysis, and administrative delays in regulatory documentation. A major reason may be the reliance on true clinical endpoints, which often takes considerable time to observe. The FDA typically evaluates new drugs by comparing true clinical endpoints across treatment groups. For example, overall survival is a true clinical endpoint that directly reflects patient clinical benefit and is widely regarded as the golden standard in clinical trials. However, measuring overall survival generally requires long-term follow-up and extends trial timelines.

To accelerate the drug approval process, researchers have explored the use of early clinical endpoints, which is known as surrogate endpoints. These are outcomes that can be measured much earlier or more conveniently in a study compared to true endpoints. An example of surrogate outcome is tumor response, which is a categorical outcome that can be observed earlier than time-to-event outcomes. A validated surrogate outcome can also be used to determine whether a new drug is eligible for approval. However, reliance on non-validated surrogate endpoints may lead to approval of drugs that provide little or no true clinical benefit. According to FDA guidelines, validation of a surrogate endpoint requires demonstrating two key associations: individual-level association (I-association) and trial-level association (T-association) (Shyr and Shyr, 2020). I-association refers to whether the surrogate is associated with the true endpoint within individuals (e.g., whether tumor response is associated with overall survival). T-association refers to whether the treatment effects on the surrogate endpoint align with the treatment effects on the true endpoint across trials. For instance, if the surrogate is binary, treatment effect might be measured using an odds ratio; if the true endpoint is survival, treatment effect might be measured using a hazard ratio. Thus, evaluating T-association indicates assessing how the odds ratio and hazard ratio relate.

Many existing studies have evaluated surrogate endpoints by reporting only the I-association. For instance, Yee et al. (2020) suggested that pathologic complete response was a surrogate for survival endpoints based on strong I-association. However, this evidence alone is not sufficient due to the ignorance of T-association. A key reason for overlooking T-association is the lack of statistical methods in modeling two random variables. When evaluating T-association, we are regressing one treatment effect against another, which are both estimated from data and are thus random variables. In this research, we aim to address this gap by developing a frequentist approach as well as a

Bayesian approach that enable rigorous assessment of T-association.

2 Method

2.1 Model Setup

Let Y_{i1} and Y_{i2} denote the treatment effects on the surrogate and true endpoint, respectively, after applying the necessary mathematical transformations (e.g., logarithmic scale). By standard large sample theory, such transformed treatment effect estimators are asymptotically normally distributed. For illustration, if the surrogate endpoint is binary, the treatment effect may be summarized by an odds ratio (OR) from a logistic regression model; if the true endpoint is time-to-event, the treatment effect may be summarized by a hazard ratio (HR) from a Cox proportional hazard model. In this setting, the observed values from study i can be expressed as $Y_{i1} = \log(\text{OR})$ and $Y_{i2} = \log(\text{HR})$. Let s_{i1}^2 and s_{i2}^2 denote the within-study variances of the estimated treatment effect for the surrogate and true endpoints, respectively. Under large sample theory, we have

$$Y_{i1} \sim N\left(\beta_1, s_{i1}^2\right) \quad \text{and} \quad Y_{i2} \sim N\left(\beta_2, s_{i2}^2\right),$$

where β_1 and β_2 represent the corresponding true underlying treatment effect.

To account for the correlation between the two treatment effects, we employ a multivariate meta-analysis model rather than modeling each treatment effect separately. Multivariate meta-analysis allows for synthesis of multiple correlated outcomes. Conventional multivariate meta-analysis partitions both the variance and the correlation into between- and within-study components; however, the within-study correlations are rarely available in practice. To address this limitation, Riley et al. (2008) proposed an alternative bivariate random-effects meta-analysis (BRMA) model, which is formulated as follows:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \Phi_i\right), \Phi_i = \begin{pmatrix} \psi_1^2 + s_{i1}^2 & \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)} \\ \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)} & \psi_2^2 + s_{i2}^2 \end{pmatrix}$$

where ψ_1^2 and ψ_2^2 are additional variation beyond s_{i1}^2 and s_{i2}^2 of the two treatment effects, respectively; and ρ is the overall correlation coefficient. ψ_1^2 and ψ_2^2 are similar to the concept of between-study variances but are not exactly equivalent because the overall correlation is not partitioned and the model does not have a fully hierarchical structure. Since the marginal distribution of the transformed treatment effects Y_{i1} and Y_{i2} are approximately Normal and they can be correlated within a study, a bivariate normal model provides a reasonable framework. Accordingly, we propose to model the pair of treatment effects (Y_{i1}, Y_{i2}) using the BRMA model, in which the correlation parameter ρ quantifies the level of T-association and serves as the key evaluation metric in our study.

This model has two advantages. First, the variance is partitioned into between- and within-study components, allowing known within-study variances to appropriately weight each study. Second, the correlation between treatment effect estimates is captured by a single parameter, which is a practical advantage because within-study correlations are

rarely reported. To estimate the parameters of the alternative BRMA model, we develop both frequentist and Bayesian approaches and construct confidence intervals to facilitate statistical inferences.

2.2 Frequentist Approach

Under the frequentist framework, maximum likelihood (ML) and restricted maximum likelihood (REML) are commonly used to estimate model parameters. In the following, we formulate the objective functions used to obtain the ML and REML estimators and discuss the construction of their corresponding confidence intervals.

ML

For the BRMA model, the likelihood function is:

$$L(\boldsymbol{\theta}|y_{i1}, y_{i2}) = \left(\frac{1}{2\pi\sqrt{1-\rho^2}}\right)^n \prod_i \left[(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)\right]^{-1/2} \exp\left\{-\frac{1}{2}\sum_i q_i\right\},$$

where $q_i = \frac{1}{1-\rho^2}\left[\frac{(y_{i1}-\beta_1)^2}{\psi_1^2+s_{i1}^2} - 2\rho\left(\frac{(y_{i1}-\beta_1)}{\sqrt{\psi_1^2+s_{i1}^2}}\frac{(y_{i2}-\beta_2)}{\sqrt{\psi_2^2+s_{i2}^2}}\right) + \frac{(y_{i2}-\beta_2)^2}{\psi_2^2+s_{i2}^2}\right]$. The ML estimators are obtained by maximizing the corresponding log-likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\theta}|y_{i1}, y_{i2}) &= \log L(\boldsymbol{\theta}|y_{i1}, y_{i2}) \\ &= -n\log(2\pi) - \frac{1}{2}\sum_i \log(\psi_1^2 + s_{i1}^2) - \frac{1}{2}\sum_i \log(\psi_2^2 + s_{i2}^2) - \frac{n}{2}\log(1-\rho^2) - \frac{1}{2}\sum_i q_i.\end{aligned}$$

The score functions are derived in the Appendix, and numerical methods are utilized to iteratively solve for the ML estimates.

REML

ML estimators for variance components are biased in finite samples. To address this, REML transforms the data to remove the influence of mean effects. The restricted likelihood is then maximized with respect to the remaining parameters based on the transformed data. Once the variance components are estimated, the mean effects are estimated via generalized least squares. In particular, the log restricted likelihood function is:

$$\begin{aligned}\ell_R(\theta|y_{i1}, y_{i2}) = -\frac{1}{2}\Big[&(n-k)\log(2\pi) - n\log|\boldsymbol{X}'\boldsymbol{X}| + \log|\boldsymbol{\Phi}| + \log\left|\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right| \\ &+ (\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Phi}^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{\beta})\Big],\end{aligned}$$

where n is the number of studies, $k = 2$ since there are two endpoints, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, and $\boldsymbol{\Phi}$ is a block-diagonal matrix with the covariance Φ_i for each study on the diagonal.

Inference

For both ML and REML estimators, the asymptotic normality of maximum likelihood estimators generally holds, and their variance can be estimated by the observed information matrix. However, when the number of studies is limited, these large-sample approximations may not hold. To address this, we employ a parametric bootstrap approach to construct robust confidence intervals for the parameter estimates. Numerical methods are also utilized to iteratively solve for the REML estimates.

2.3 Bayesian Approach

Bayesian methods are often considered more robust in small-sample settings. In real-world applications, the number of trials available to evaluate T-association may be limited, making Bayesian inference a useful alternative. Flat priors are assigned to all model parameters, and several alternative prior distributions are examined to evaluate robustness. As the posterior distribution lacks a closed-form, we rely on Markov Chain Monte Carlo (MCMC) methods to draw samples and approximate the posterior.

When performing MCMC sampling, instead of using the traditional Metropolis-Hastings algorithm, we adopt Hamiltonian Monte Carlo (HMC), which is a more modern and efficient technique. HMC leverages gradient information from the posterior to propose distant, high-probability values, improving sampling efficiency compared to random-walk-based methods. This is particularly advantageous when parameters are strongly correlated, as in our setting.

2.4 Model extension

Clinical trial designs may vary in several aspects, such as the sample size or treatment regimen, which can introduce heterogeneity across trials. To account for this heterogeneity, we propose extending the model to incorporate trial-level covariates. The extended model is given by:

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{e}_i, \quad \boldsymbol{e}_i \sim N\left(\mathbf{0}, \Phi_i\right),$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \textit{and} \quad \Phi_i = \begin{pmatrix} \psi_1^2 + s_{i1}^2 & \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)} \\ \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)} & \psi_2^2 + s_{i2}^2 \end{pmatrix}.$$

The derivations for the estimators similar to those in the original model, with the mean replaced by $(\boldsymbol{X\beta_1}, \boldsymbol{X\beta_2})$ instead of $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$. Therefore, the detailed derivations are omitted here.

3 Numerical Studies

3.1 Simulation studies

To evaluate the performance of the proposed methods, we perform simulation studies. We generated data from the alternative BRMA model. The parameters are set at $\boldsymbol{\beta} = (0.4, 0.7)$, $\psi_1^2 = 0.25$, $\psi_2^2 = 0.36$, and $\rho = 0.6$. In real cases, s_{i1}^2 and s_{i2}^2 are known; to mimic this scenario, we generated s_{i1}^2 and s_{i2}^2 randomly from a uniform distribution with range 0.01 to 0.09. To evaluate the effect of sample size, we considered different settings of n, namely $20, 8$ and 6. For each setting, 1000 datasets were generated. We estimated each of the data sets using the two frequentist approaches and the Bayesian approach.

Frequentist approach

For the frequentist approach, we report bias, empirical standard error ($\mathrm{SE_{em}}$), sample average of the theoretical standard error ($\mathrm{SE_{th}}$), and coverage probability of the 95% confidence intervals constructed under the normal assumption ($\mathrm{CP_{th}}$) for each parameter. We also performed 1,000 parametric bootstrap replications to construct 95% bootstrapped confidence intervals and reported the corresponding coverage probabilities ($\mathrm{CP_b}$).

The results are shown in Table 1. All biases are small, and bias generally reduces as the sample size increases. For variance components, REML shows smaller bias compared to ML. $\mathrm{SE_{em}}$ and $\mathrm{SE_{th}}$ are well-aligned for a sample size of 20, making $\mathrm{CP_{th}}$ reasonable. For smaller sample sizes, $\mathrm{SE_{th}}$ tends to underestimate variability, which is also reflected in the low coverage probability of $\mathrm{CP_{th}}$. Therefore, we recommend using the parametric bootstrap method to construct confidence intervals is number of trials drop below 10, since $\mathrm{CP_b}$ is generally more robust than $\mathrm{CP_{th}}$ in small sample settings. Overall, REML performs better than ML in terms of bias and coverage probabilities..

Bayesian approach

For the Bayesian approach, we assigned the following flat priors to each parameter: Normal(0, 2.5^2) for β_1 and β_2, Inverse-Gamma(0.01, 0.01) for ψ_1^2 and ψ_2^2, and Uniform(-1, 1) for ρ. We then used the HMC method to draw samples and approximate the posterior distribution.

To evaluate the simulation results, we reported the bias, the mean of the posterior standard deviations (SD_p), and the empirical coverage (EC). The results are shown in Table 2. The bias is slightly larger than in the frequentist approach, likely due to the use of flat priors. However, the empirical coverage is closer to 95%, indicating that the Bayesian approach is generally more robust than the frequentist approach.

3.2 Data analysis

To demonstrate our proposed method, we collected data from 14 non-small cell lung cancer trials comparing chemotherapy and immunotherapy, each reporting hazard ratios for progression-free survival(PFS) and overall survival. While PFS is not an ideal surrogate

Table 1: Simulation results for parameter estimation using the frequentist approach.

n	Method	Param	Bias	$\mathrm{SE}_{\mathrm{em}}$	$\mathrm{SE}_{\mathrm{th}}$	$\mathrm{CP}_{\mathrm{th}}$	CP_{b}
20	ML	β_1	0.0003	0.1267	0.1198	93.1	93.2
		β_2	0.0016	0.1404	0.1363	92.3	92.4
		ψ_1^2	-0.0032	0.0913	0.0962	91.7	93.8
		ψ_2^2	-0.0221	0.1235	0.1210	87.5	87.9
		ρ	-0.0130	0.1524	0.1440	91.1	94.1
	REML	β_1	0.0007	0.1266	0.1209	93.7	92.9
		β_2	0.0019	0.1407	0.1396	92.6	92.9
		ψ_1^2	0.0041	0.0955	0.0972	91.9	93.6
		ψ_2^2	-0.0027	0.1293	0.1296	90.5	92.3
		ρ	-0.0130	0.1519	0.1449	91.0	94.7
8	ML	β_1	0.0054	0.1859	0.1878	95.2	96.3
		β_2	0.0043	0.2143	0.2108	93.1	95.2
		ψ_1^2	-0.0093	0.1382	0.1575	87.7	91.7
		ψ_2^2	-0.0281	0.1871	0.1912	82.6	86.1
		ρ	-0.0121	0.2604	0.2147	82.2	90.2
	REML	β_1	0.0061	0.1846	0.1933	95.0	95.1
		β_2	0.0049	0.2143	0.2239	95.2	94.5
		ψ_1^2	0.0191	0.1574	0.1702	90.1	93.6
		ψ_2^2	0.0245	0.2143	0.2283	89.1	92.6
		ρ	-0.0114	0.2574	0.2219	83.4	91.5
6	ML	β_1	-0.0083	0.1990	0.2134	96.0	97.9
		β_2	0.0002	0.2383	0.2506	95.1	97.3
		ψ_1^2	-0.0125	0.1631	0.1621	79.9	88.2
		ψ_2^2	-0.0228	0.2209	0.2325	82.5	87.7
		ρ	-0.0176	0.3025	0.2212	76.6	87.0
	REML	β_1	-0.0086	0.1970	0.2345	97.5	97.6
		β_2	0.0006	0.2387	0.2721	96.5	96.8
		ψ_1^2	0.0599	0.1957	0.2253	92.8	96.2
		ψ_2^2	0.0565	0.2658	0.3027	89.7	94.3
		ρ	-0.0177	0.3026	0.2543	80.5	93.7

Table 2: Simulation results for parameter estimation using the Bayesian approach.

n	Param	Bias	Posterior SD	Empirical Coverage
20	β_1	0.0008	0.1256	94.9
	β_2	-0.0034	0.1485	94.8
	ψ_1^2	0.0275	0.1161	95.4
	ψ_2^2	0.0458	0.1621	95.8
	ρ	-0.0776	0.1553	95.0
8	β_1	-0.0099	0.2095	95.0
	β_2	-0.0180	0.2447	93.6
	ψ_1^2	0.0837	0.2881	95.7
	ψ_2^2	0.1135	0.3931	95.7
	ρ	-0.1836	0.2577	93.7
6	β_1	0.0045	0.2579	94.3
	β_2	-0.0068	0.3005	93.2
	ψ_1^2	0.1576	0.4969	95.0
	ψ_2^2	0.2156	0.6792	94.4
	ρ	-0.2115	0.2982	94.6

since it is a time-to-event outcome and also takes a long time to observe, it is commonly reported and serves well for demonstration.

In Figure 1, we visualized the uncertainty of the treatment effect estimates using ellipses. Since within-trial correlations are not available, the ellipses do not reflect any potential association between the two estimates, and therefore all appear upright.

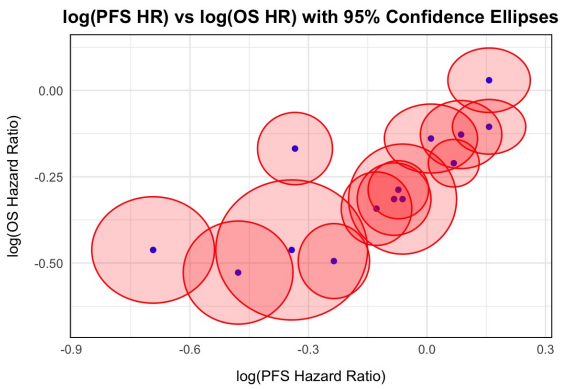

Figure 1: Visualization of both treatment effect estimates for 14 trials.

The estimated correlations (ρ) between log(PFS) and log(OS) are presented in Table 3. Overall, the correlation between the two treatment effect estimates is approximately 0.7, indicating a high correlation. These findings provide supportive evidence that can be submitted to the FDA for consideration when evaluating whether PFS is a valid surrogate for OS in terms of T-association.

Table 3: Estimates of ρ for real dataset.

Method	Estimates (95% CI)
ML	0.76 (0.51-0.90)
REML	0.76 (0.48-0.92)
Bayesian	0.70 (0.29-0.88)

4 Discussion

According to FDA guidelines, validation of a surrogate endpoint requires assessment of both I-association and T-association. In practice, however, only I-association is commonly evaluated, which is not sufficient to establish surrogacy. Evaluating T-association is also necessary but is often overlooked due to statistical challenges. In this research, we proposed both frequentist and Bayesian approaches to evaluate T-association. We assumed that the two treatment effect estimates follow a bivariate Normal model proposed by Riley et al. (2008). To account for trial-level heterogeneity, the model is further extended to incorporate trial-level covariates. Simulation results suggest that REML performs better than ML, while the Bayesian approach provides more robust inference than the frequentist approach.

There are some limitations in this study. The bivariate Normal assumption of the treatment effect estimates may be restrictive, particularly when treatment effect estimates are skewed or heavy-tailed. We introduced diagnostic methods to assess this distributional assumption; however, with limited sample sizes, the diagnostic results may be inconclusive or unstable. Heavy-tailed or more flexible models could be considered as future work for more general application. Another limitation is that only a small number of trial-level covariates can be incorporated to adjust for heterogeneity due to limited trial sizes, making it important to identify or design clinical trials with similar characteristics to reduce potential bias in the estimates. Overall, this study provides a practical framework for evaluating the T-association of a surrogate endpoint and can serve as the statistical foundation for future FDA accelerated drug approval processes.

References

Brown, D. G., Wobst, H. J., Kapoor, A., Kenna, L. A., and Southall, N. (2022). Clinical development times for innovative drugs. *Nat Rev Drug Discov*, 21(11):793–794.

Riley, R. D., Thompson, J. R., and Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9(1):172–186.

Shyr, Y. and Shyr, D. (2020). What constitutes a valid surrogate end point in cancer clinical trials? *JAMA oncology*, 6(9):1334–1335.

Stewart, D. J., Stewart, A. A., Wheatley-Price, P., Batist, G., Kantarjian, H. M., Schiller, J., Clemons, M., Bradford, J.-P., Gillespie, L., and Kurzrock, R. (2018). The importance of greater speed in drug development for advanced malignancies. *Cancer medicine*, 7(5):1824–1836.

Yee, D., DeMichele, A. M., Yau, C., Isaacs, C., Symmans, W. F., Albain, K. S., Chen, Y.-Y., Krings, G., Wei, S., Harada, S., et al. (2020). Association of event-free and distant recurrence–free survival with individual-level pathologic complete response in neoadjuvant treatment of stages 2 and 3 breast cancer: three-year follow-up analysis for the i-spy2 adaptively randomized clinical trial. *JAMA oncology*, 6(9):1355–1362.