Biostatistics (2014), **0**, 0, *pp.* 1–9
doi:10.1093/biostatistics/Bios˙8cat˙supp˙REV2

Treatment Selections using Risk-Benefit Profiles Based on Data from Comparative Randomized Clinical Trials with Multiple Endpoints: Supplementary Materials

BRIAN CLAGGETT*

Division of Cardiovascular Medicine, Harvard Medical School, Boston, Massachusetts, U.S.A.

bclaggett@partners.org

LU TIAN

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, U.S.A.

DAVIDE CASTAGNO

Division of Cardiology, Department of Medical Sciences, University of Turin, Italy

LEE-JEN WEI

Department of Biostatistics, Harvard University, Boston, Massachusetts, U.S.A.

APPENDIX

For all inference using large sample approximations, we employ perturbation-resampling procedures using 1000 realizations from the standard exponential distribution. Details are provided below.

Appendix A: Construction of Confidence Intervals for Two-Sample Inference

Let $\{B_{ij} : i = 1, 2; j = 1, \ldots, n_i\}$ be independent random samples from a strictly positive distribution with mean and variance equal to one. The cumulative cell probability γ_{ik} can be estimated by

$$\hat{\gamma}_{ik} = \sum_{j=1}^{n_i} \mathbb{W}_{ij} I(\epsilon_{ij} \leqslant k) / \sum_{j=1}^{n_i} \mathbb{W}_{ij} \quad \text{or} \quad \sum_{j=1}^{n_i} \widetilde{\mathbb{W}}_{ijk} I(\epsilon_{ij} \leqslant k) / \sum_{j=1}^{n_i} \widetilde{\mathbb{W}}_{ijk}.$$

Let γ^*_{ik} be the corresponding perturbed version of $\hat{\gamma}_{ik}$ with

$$\gamma^*_{ik} = \sum_{j=1}^{n_i} \mathbb{W}^*_{ij} I(\epsilon_{ij} \leqslant k) / \sum_{j=1}^{n_i} \mathbb{W}^*_{ij} \quad \text{or} \quad \sum_{j=1}^{n_i} \widetilde{\mathbb{W}}^*_{ijk} I(\epsilon_{ij} \leqslant k) / \sum_{j=1}^{n_i} \widetilde{\mathbb{W}}^*_{ijk}, \tag{A.1}$$

where

$$\mathbb{W}^*_{ij} = B_{ij} \frac{I(T_{ij} \wedge t_0 \leqslant C_{ij})}{\hat{G}^*_i(T_{ij} \wedge t_0)} \quad \text{and} \quad \widetilde{\mathbb{W}}^*_{ijk} = B_{ij} \frac{I(T_{ijk} \leqslant C_{ij})}{\hat{G}^*_{ik}(T_{ijk})}.$$

Here, both $\hat{G}^*_i(\cdot)$ and $\hat{G}^*_{ik}(\cdot)$ are the perturbed estimators for the survival function $G_i(\cdot)$:

$$\hat{G}^*_i(t) = \exp\left[-\sum_{j=1}^{n_i} \int_0^t \frac{B_{ij} d\{I(u \leqslant T_{ij} < C_{ij})\}}{\sum_{l=1}^{n_i} B_{il} I(X_{il} \geqslant u)} \right] \tag{A.2}$$

and

$$\hat{G}^*_{ik}(t) = \exp\left[-\sum_{j=1}^{n_i} \int_0^t \frac{B_{ij} d\{I(u \leqslant T_{ijk} < C_{ij})\}}{\sum_{l=1}^{n_i} B_{il} I(T_{ijk} \wedge C_{ij} \geqslant u)} \right] \tag{A.3}$$

Furthermore, let β^* be the maximizer of the perturbed version of the weighted log-likelihood function (2.4):

$$\sum_{k=1}^{K-1} \sum_{ij} \widetilde{\mathbb{W}}^*_{ijk} [I(\epsilon_{ij} \leqslant k) log\{g^{-1}(\alpha_k - \beta\tau_{ij})\} + I(\epsilon_{ij} > k) \log\{1 - g^{-1}(\alpha_k - \beta\tau_{ij})\}]. \tag{A.4}$$

The limiting distribution, conditional on the data, of

$$(n_1 + n_2)^{1/2} (\beta^* - \hat{\beta}), \tag{A.5}$$

is normal with mean 0 and variance $\hat{\sigma}^2_b$, which is a consistent estimator of σ^2_b, the variance associated with the distribution $(n_1+n_2)^{1/2}(\hat{\beta}-\beta)$. Thus, the empirical variance of the perturbed

estimates β^* can be used to estimate the standard error associated with $\hat{\beta}$ (Zheng *and others*, 2006; Uno *and others*, 2007; Li *and others*, 2011).

Denote $\mathbf{\Gamma}^* = \boldsymbol{\gamma}_2^* - \boldsymbol{\gamma}_1^*$, where $\boldsymbol{\gamma}_i^* = \{\gamma_{i1}^*, \cdots, \gamma_{iK}^*\}'$. Using the arguments by Cai *and others* (2010), the limiting distribution, conditional on the target data set, of

$$(n_1 + n_2)^{1/2}(\mathbf{\Gamma}^* - \hat{\mathbf{\Gamma}}), \tag{A.6}$$

is multivariate normal with mean zero and covariance matrix $\hat{\mathbf{\Sigma}}$ which is a consistent estimator of $\mathbf{\Sigma}$, the covariance matrix associated with the distribution $(n_1+n_2)^{1/2}(\hat{\mathbf{\Gamma}}-\mathbf{\Gamma})$. Thus, the resulting sample covariance matrix based on those perturbed estimates $\mathbf{\Gamma}^*$, say, $\tilde{\mathbf{\Sigma}}$, is a valid estimator of $\mathbf{\Sigma}$. A two-sided confidence interval for the two-sample risk difference Γ_k is then given by

$$\hat{\Gamma}_k \pm z_{(1-\alpha/2)}(n_1 + n_2)^{-1/2}\tilde{\sigma}_k, \tag{A.7}$$

where $\tilde{\sigma}_k^2$ is the kth diagonal element of $\tilde{\mathbf{\Sigma}}$. Furthermore, one may use a similar approach for making inference on $\hat{D}$ by perturbed $D^* = \sum_{k=2}^{K} \pi_{1k}^* \gamma_{2(k-1)}^* - \pi_{2k}^* \gamma_{1(k-1)}^*$, where $\pi_{ik}^* = \gamma_{ik}^* - \gamma_{i(k-1)}^*$.

Appendix B: Construction of Confidence Intervals and Bands for Stratified Inference

For personalized medicine, we let $\gamma_{ik}^*(s)$ be the perturbed version of $\hat{\gamma}_{ik}(s)$ with

$$\gamma_{ik}^*(s) = \left\{ \sum_{j=1}^{n_i^*} \widetilde{\mathbb{W}}_{ijk} I(\epsilon_{ij} \leqslant k) K_{h_i}(V_{ij} - s) \right\} \Big/ \left\{ \sum_{j=1}^{n_i^*} \widetilde{\mathbb{W}}_{ijk} K_{h_i}(V_{ij} - s) \right\}. \tag{A.8}$$

and $\pi_{ik}^*(s) = \gamma_{ik}^*(s) - \gamma_{i(k-1)}^*(s)$. Using identical arguments to those above, we denote $\mathbf{\Gamma}^*(s) = \boldsymbol{\gamma}_2^*(s) - \boldsymbol{\gamma}_1^*(s)$, where $\boldsymbol{\gamma}_i^*(s) = \{\gamma_{i1}^*(s), \cdots, \gamma_{iK}^*(s)\}'$, and can show that the distribution for

$$(n_1^* h_1 + n_2^* h_2)^{1/2}\{\mathbf{\Gamma}^*(s) - \hat{\mathbf{\Gamma}}(s)\}, \tag{A.9}$$

conditional on the observed data, is multivariate normal and asymptotically equivalent to that of $(n_1^* h_1 + n_2^* h_2)^{1/2}\{\hat{\mathbf{\Gamma}}(s) - \mathbf{\Gamma}(s)\}$. Therefore, the point-wise confidence interval for $\Gamma(s)$ can be constructed using generated $\Gamma^*(s)$ as in (A.5).

To construct a $(1-\alpha)$ simultaneous confidence band for $\Gamma_k(s)$ over the pre-specified interval $\mathcal{S}$, we cannot use the conventional method based on the sup-statistic,

$$\sup_{s\in\mathcal{S}} \tilde{\sigma}_k^{-1}(s)|(n_1^* h_1 + n_2^* h_2)^{1/2}\{\hat{\Gamma}_k(s) - \Gamma_k(s)\}| \tag{A.10}$$

due to the fact that as a process in s, $(n_1^* h_1 + n_2^* h_2)^{1/2}\{\hat{\Gamma}_k(s) - \Gamma_k(s)\}$ does not converge weakly to a tight process. On the other hand, one may utilize the strong approximation argument given in Bickel and Rosenblatt (1973) to show that an appropriately transformed sup of $\hat{\Gamma}_k(s) - \Gamma_k(s)$ converges to a proper random variable. In practice, to construct a confidence band, we can first find a critical value b_α such that

$$\text{pr}\left(\sup_{s\in\mathcal{S}} |\Gamma_k^*(s) - \hat{\Gamma}_k(s)|/\{(n_1^* h_1 + n_2^* h_2)^{-1/2}\tilde{\sigma}_k(s)\} > b_\alpha \,\middle|\, \text{observed data}\right) \approx \alpha. \tag{A.11}$$

Then the confidence band for $\Gamma_k(s) : s \in \mathcal{S}$ is given by

$$\hat{\Gamma}_k(s) \pm b_\alpha (n_1^* h_1 + n_2^* h_2)^{-1/2}\tilde{\sigma}_k(s). \tag{A.12}$$

Similar arguments can be used for the construction of the confidence band for $E(s) : s \in \mathcal{S}$.

Appendix C: BEST treatment differences with respect to specific outcome thresholds

In Figure A.3, we show the smoothed treatment effect estimates with respect to each definition of treatment success, i.e. $(\epsilon \leqslant k)$, in the part B data set, conditional on the model-based treatment effect score $\hat{d}(u)$ derived from the part A data. Treatment responses appear to be most predictable with respect to $k = 1, 2$, and 3.

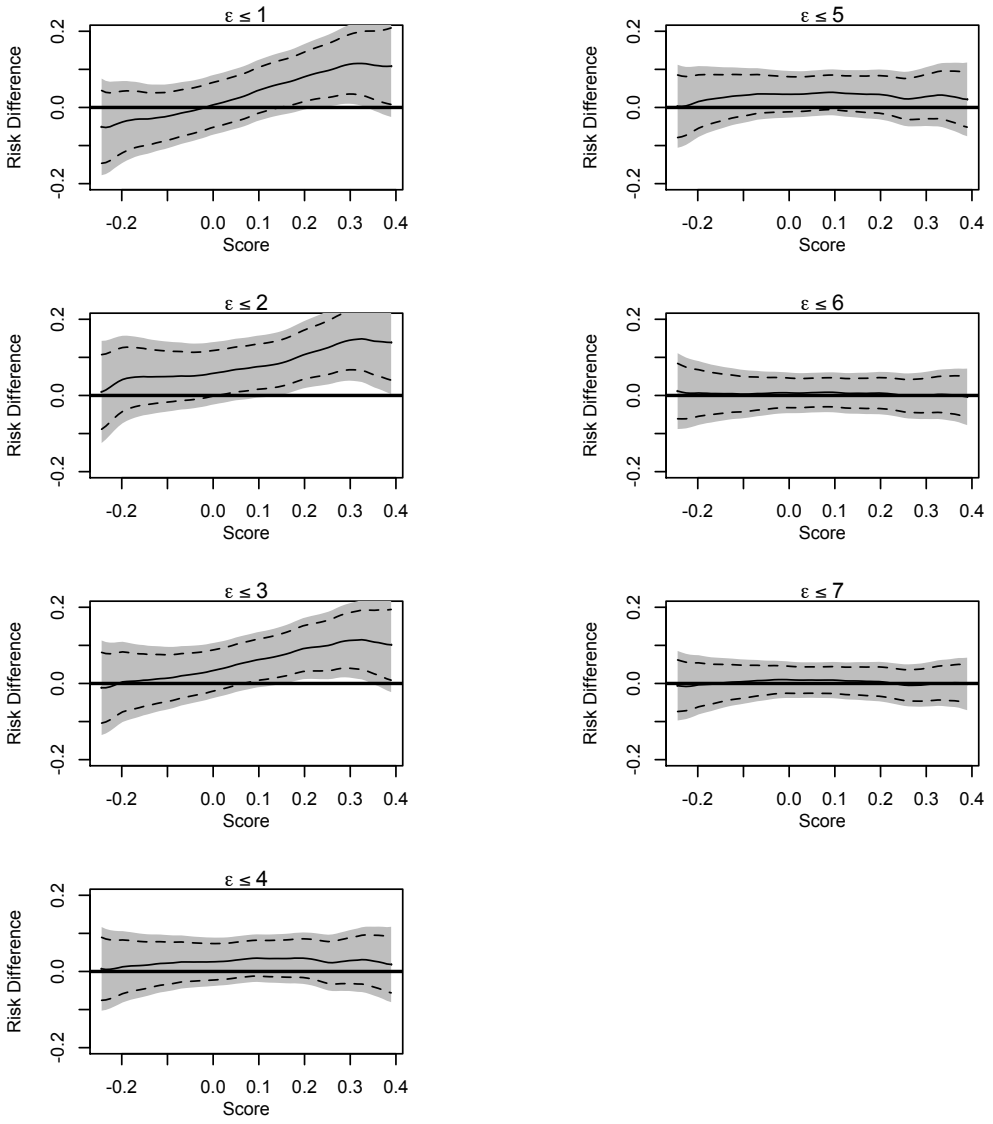

Fig. A.1. BEST target treatment differences (active minus placebo) using treatment selection score derive from the model in Table **??**. Solid curve represents point estimates, with 0.95 pointwise and simultaneous confidence intervals denoted by dashed lines and shaded region, respectively.

Appendix D: Results of 36-month analysis

Using patient outcomes evaluated at $t_0 = 36$ months, the overall treatment effect is estimated to be $\hat{D} = 0.077$ with standard error estimate of 0.032. The estimated distributions of patient

Table A.1. Estimated distribution functions for control and treated groups with BEST data with $t_0 = 36$ months

	Control ($\hat{\gamma_1}$)		**Treated ($\hat{\gamma_2}$)**		**Contrast ($\hat{\Gamma}$)**	
Outcome Category	n	pr($\epsilon \leqslant k$)	n	pr($\epsilon \leqslant k$)	Est	SE
1	94	0.19	112	0.23	+0.04	0.02
2	80	0.36	101	0.44	+0.08	0.03
3	53	0.46	43	0.54	+0.08	0.03
4	67	0.60	48	0.64	+0.03	0.02
5	31	0.64	41	0.68	+0.04	0.02
6	205	0.86	171	0.87	+0.01	0.01
7	24	0.88	22	0.88	+0.01	0.01
8	163	1.00	153	1.00	-	-
(censored)	636	-	663	-	-	-

outcomes and associated contrast measures are shown in Table A.1. The patient-specific treatment differences are shown in Figure A.2.

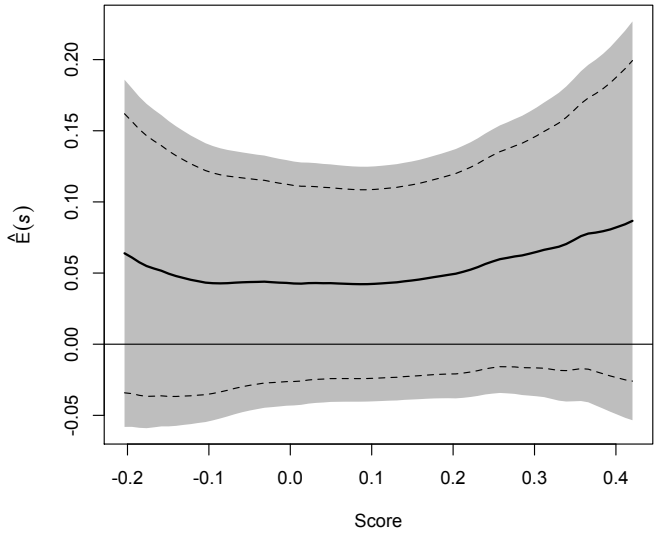

Fig. A.2. Estimated BEST treatment effect $\hat{E}(s)$ using treatment selection score presented in Table **??** with outcomes obtained at $t_0 = 36$ months. Solid curve represents point estimates, with 0.95 pointwise and simultaneous confidence intervals denoted by dashed lines and shaded region, respectively.

Appendix E: Simulation Details: Comparison of ordinal vs binomial models and model fit information

In Figure A.3, we provide the average curves, $\{\hat{D}^*(q)-\hat{D}\}$, resulting from the ordinal and binomial logistic regression models, using the complementary log-log link and "partial information" weights $\widetilde{\mathbb{W}}_{ijk}$, applied to 200 simulated data sets generated under simulation scenario #1. Ten-fold cross-validation was performed within each data set, and the concordance statistics $\hat{\mathbb{C}}$ are calculated as $\hat{\mathbb{C}} = \int_0^1 (1-q)\{\hat{D}^*(q) - \hat{D}\}dq$, as described in Section 3.

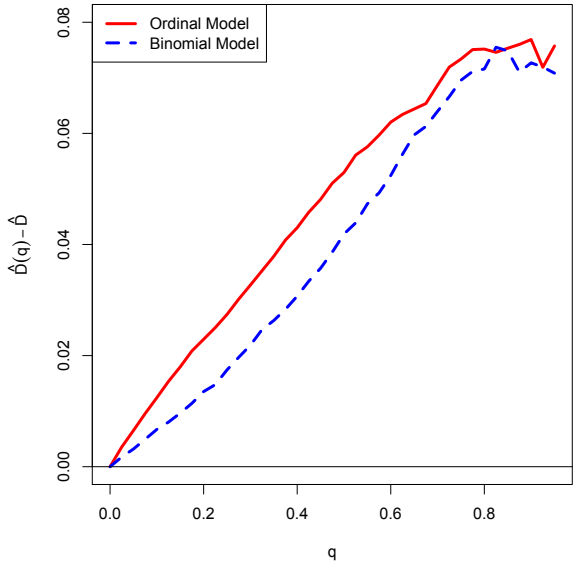

Fig. A.3. Comparison of ordinal vs binomial models (simulation scenario #1) for the purpose of stratifying patients according to predicted treatment response. $\hat{D}(q)$: Overall treatment difference among patients with top $100(1-q)\%$ of scores.

Although we cannot obtain the explicit expression of $D(u)$ as a function of u, we can obtain $D(u)$ for each patient through Monte Carlo simulation. For any given patient with U, we can then compare his/her $D(u)$ and the approximate $E(s)$. We note that with a continuous scoring

system, no two patients will necessarily have identical values of s. However, we can approximate $E(s)$ by estimating the true treatment effect over all patients who belong to the same decile of the working model score, s. For illustration, we choose three patients from the data set of BEST, representing patients with negative, neutral, and positive values of $D(u)$, where D represents the net probability of treatment benefit (i.e. probability of benefit from active therapy minus probability of benefit from placebo). The patients' covariate profiles are listed in the following table. On the bottom of the table, we include $D(u)$ and $E(s)$.

Table A.2. Representative patient profiles from Simulation Setting #1

Example Patient #	1	2	3
Age	56	72	30
Gender	Female	Male	Male
LVEF	22	34	24
eGFR Category	<45	60-75	> 75
SBP	148	100	128
NYHA class	3	3	3
Obesity	No	Yes	No
Ever smoked	Yes	Yes	Yes
Heart Rate	68	88	110
Hypertension	Yes	Yes	Yes
Diabetes	Yes	No	No
Ischemic HF	No	Yes	No
Atrial Fibrillation	No	No	No
Race	Black	White	White
$D_0(U)$	-0.07	0	0.08
E(s)	-0.06	0	0.06

We further evaluated the model performance in simulation setting #1. Using this approximation, we find overall coverage of 91% of the true $E(s)$ across replications. To assess classification accuracy, we compare the observed MSE [$\hat{E}(s)$ vs $E(s)$] against that from a null model [$\bar{E}(s)$ vs $E(s)$]. We estimate the MSE based on stratification of patients via the working models to be 0.0034, compared to 0.0051 for the MSE from the null model. The ratio of these values is approximately 0.67, suggesting a "pseudo-R^2 " value of 0.33.

In terms of classification accuracy, we may utilize the standard AUC metric to quantify the accuracy with which the working model in a given data set can identify patients who would truly benefit from treatment (i.e. $D(U) > 0$). We find that the median AUC within a single replicate data set is 0.73 [IQR: 0.69-0.77], while the usage of each patient's limiting score (averaged over all replications) yielded a much higher AUC = 0.88. This difference in AUC is expected, as the former reflects both the uncertainty in fitting the working model and the quality of the working model, the latter is determined only by the quality of the working model.

References

Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1071–1095.

Cai, T., Tian, L., Uno, H., Solomon, S.D. and Wei, LJ. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* **97**(2), 389–404.

Li, Y., Tian, L. and Wei, L.J. (2011). Estimating subject-specific dependent competing risk profile with censored event time observations. *Biometrics* **67**(2), 427–435.

Uno, H., Cai, T., Tian, L. and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *JASA* **102**, 527–537.

Zheng, Y., Cai, T. and Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62**(1), 279–287.