BIOMETRICS 00, 1–5 DOI: 10.1111/j.1541-0420.2005.00454.x
December 0000

Web-based Supplementary Materials for "Combining Biomarkers to Optimize Patient Treatment Recommendations"

by Chaeryon Kang, Holly Janes, and Ying Huang

Vaccine and Infectious Disease Division and Public Health Sciences Division,

Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A. Choice of tuning parameters

A-1. Choice of the maximum number of iterations, M_{max}, and weight, $\widetilde{w}\{\Delta(Y)\}$

There are two tuning parameters that need specification when implementing the boosting method: the weight function, $\widetilde{w}\{\Delta(Y)\}$, and the maximum number of iterations, $\mathrm{M_{max}}$. Choosing $\mathrm{M_{max}}$ is similar to choosing the number of base-models in any ensemble method that combines multiple base-models (Opitz and Maclin (1999); Assareh et al. (2008)). Typically a larger number of base-models yields improved model performance, up until some $\mathrm{M_0}$ beyond which no improvement and potentially even deterioration in performance is observed. The best weight function and optimal $\mathrm{M_{max}}$ are not known in practice, and so we recommend investigating these choices using a separate data set that is not used for fitting or evaluating model performance, or using cross-validation (CV).

A-1-1. Impact of choice of M_{max} and $\widetilde{w}\{\Delta(Y)\}$ in simulations. In our simulation study, we set $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{-\frac{1}{3}}$ which was the best-performing weight function among several for the models we considered in the sense of maximum mean θ across 1000 training data sets. In addition to $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{-\frac{1}{3}}$, we considered $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{-\frac{1}{10}}$, $\widetilde{w}\{\Delta(Y)\} = e^{-|\Delta(Y)|}$ and $\widetilde{w}\{\Delta(Y), Y\} = e^{-|\Delta(Y)|}W_{\mathrm{A}}(Y)$, where $W_{\mathrm{A}}(Y)$ is similar to the weight function used in Adaboost (Friedman et al., 2000). Specifically, $W_{\mathrm{A}}(Y) = \exp\left\{-\frac{1}{2}\log\left(\frac{1-err}{err}\right)\times(2D-1)(2\widehat{D}-1)\right\}$, where $\widehat{D} = \mathbf{1}\{\widehat{P}(D=1|T,Y) > 0.5\}$ is the outcome classification at the previous stage and $err = P(D \neq \widehat{D})$ is the error in this classification. Additional polynomial weight functions of the form $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{d}$ were also considered (data not shown). Web Table 1 compares the performance of the boosting method under different choices for the weight function for the 4 most informative simulation scenarios. The results suggest that the best-performing weight function depends on simulation scenario and working model. However, the improvement in model performance associated with using the optimal $\widetilde{w}\{\Delta(Y)\}$ was minimal.

In the simulations, $\mathrm{M}_{\mathrm{Best}}$ is what we found to be the best-performing $\mathrm{M}_{\max}$ among $\mathrm{M}_{\max} = 1, \ldots, 50$, in terms of maximizing mean θ across 1000 training data sets for each $\mathrm{M}_{\max}$. Web Figures 1, 2, 3 and 4 show that for most simulation scenarios with $n = 500$ observations, $\mathrm{M}_{\max} = 10 \sim 20$ yields near-optimal mean θ and $\mathrm{M}_{\max} = 40 \sim 50$ achieves optimal mean θ. However, as with choice of the weight function, the improvement in model performance associated with using the optimal $\mathrm{M}_{\max}$ is minimal (Web Table 1). These figures also show that $\mathrm{M}_{\mathrm{Best}}$ was also near-optimal in terms of minimizing $\mathrm{MCR}_{\mathrm{TB}}$.

A-1-2. Choosing M_{max} and $\widetilde{w}\{\Delta(Y)\}$ in practice using cross-validation. In practice, to determine the maximum number of iterations, $\mathrm{M}_{\max}$, and the best weight function, $\widetilde{w}\{\Delta(Y)\}$, we recommend K-fold cross-validation. We start with a collection of reasonable $\mathrm{M}_{\max}$, for example, $\widetilde{M}^{(1)} = \{10, 50, 100, 300, 500\}$. Using $K - 1/K$ of the data, we apply the boosting method with each of $\mathrm{M}_{\max} \in \widetilde{M}$, and estimate θ using the remaining hold-out data. We calculate $\widehat{\theta}$ as the average estimated θ over K hold-out data sets. This entire procedure is then repeated J times, where we use $J = 10$. Let $\mathrm{M}_{\max}^{(1)} = \arg\max_{\widetilde{M}^{(1)}} \widehat{\theta}$. In the second stage, we refine $\widetilde{M}^{(1)}$ further using a finer grid of possible $\mathrm{M}_{\max}$ values. For example, if $\mathrm{M}_{\max}^{(1)} = 150$, then $\widetilde{M}^{(2)} = \{100, \ldots, 130, 140, 150, 160, \ldots, 200\}$ and $\widehat{\theta}$ is calculated for each element of $\widetilde{M}^{(2)}$. The third stage refines $\widetilde{M}^{(2)}$ even further. In our analysis, we have found that 3-stages for refining $\widetilde{M}$ has been sufficient and define the best $\mathrm{M}_{\max}$ as $\mathrm{M}_{\max}^{(3)} = \arg\max_{\widetilde{M}^{(3)}} \widehat{\theta}$. In general, we recommend continuing to refine $\widetilde{M}$ until the variation in $\widehat{\theta}$ over $\widetilde{M}$ is minimal.

We recommend a similar CV procedure to determine the best weight function, $\widetilde{w}\{\Delta(Y)\}$, given a set of possible weight functions. Alternatively, one could conduct a single CV analysis, simultaneously optimizing the choice of $\mathrm{M}_{\max}$ and $\widetilde{w}\{\Delta(Y)\}$, using a grid search method. This is what we used for the breast cancer data analysis; the procedure is described in detail below.

[Web Table 1 about here.]

[Web Figure 1 about here.]

[Web Figure 2 about here.]

[Web Figure 3 about here.]

[Web Figure 4 about here.]

A-1-3. Application of the CV procedure to the breast cancer data. In the breast cancer data analysis, the best weight function and the maximum number of iterations were determined using 10 replications of 5-fold CV. We considered weight functions of the form $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^d$, where $d \in \widetilde{D}^{(1)} = \{-1.85, -1.6, -1.35, -1.1, -0.85, -0.6, -0.35, -0.1\}$. The best d and $\mathrm{M}_{\max}$ were explored using a grid search. In the first stage, we applied the boosting method for each element of $\widetilde{DM}^{(1)} = \{(d, \mathrm{M}_{\max}) : d \in \widetilde{D}^{(1)}, \mathrm{M}_{\max} \in \widetilde{M}^{(1)}\}$ to obtain $\mathrm{DM}^{(1)}_{\max} = \arg\max_{\widetilde{DM}^{(1)}} \widehat{\theta}$. In the second stage, we refined $\widetilde{DM}^{(1)}$ and performed another grid search yielding $\mathrm{DM}^{(2)}_{\max} = \arg\max_{\widetilde{DM}^{(2)}} \widehat{\theta}$. We further refined $\widetilde{DM}^{(2)}$ and performed a third grid search to obtain the best $(d, \mathrm{M}_{\max}) = \mathrm{DM}^{(3)}_{\max} = \arg\max_{\widetilde{DM}^{(3)}} \widehat{\theta}$. The resultant best weight function and maximum number of iterations are given in Web Table 2.

[Web Table 2 about here.]

A-2. Influence of the choice of maximum weight, C_M

In our simulations and data analysis we used a "weight trimming" strategy that truncates weights $\widetilde{w}\{\widetilde{\Delta}(Y_i)\}$ for subject i at a maximum weight, $\mathrm{C_M} = 500$. Weight trimming avoids highly variable estimators that result when subjects with $\widetilde{\Delta}(Y_i) \approx 0$ receive enormous weight; this strategy is commonly employed for inverse-probability weighted estimation (Potter

(1993); Cole and Hernán (2008); Lee et al. (2011)). However, under a correctly specified working model, weight trimming can reduce variance of estimation at the cost of increased bias (Cole and Hernán, 2008).

Web Table 3 shows the simulation results for the boosting method using different choices for the maximum weight; C_{M} is varied from 300 to 1000. Selected simulation scenarios with $n = 500$ observations are examined, and the linear logistic working model is used. We observe that neither the mean θ or mean $\mathrm{MCR}_{\mathrm{TB}}$ across 1000 training data sets is sensitive to the choice of C_{M} and therefore fixing $\mathrm{C}_M = 500$ appears reasonable.

[Web Table 3 about here.]

Web Appendix B. Bias-correction by bootstrap and double-bootstrap sampling

In the breast cancer data analysis, we used the bootstrap bias correction approach (Efron and Tibshirani, 1993). Briefly, given the apparent $\widehat{\theta}$ obtained using the original (training) data set, bootstrap bias estimate is $\widehat{Bias}_b(\widehat{\theta}) = \widehat{\theta} - B^{-1}\sum_{b=1}^{B}\widehat{\theta}_b$, where $\widehat{\theta}_b$ is the estimate of θ in the original training data given $\widehat{\phi}_b$ estimated using bootstrap sample b and B denotes the number of bootstrap replications. Then the bootstrap bias-corrected estimate of θ is calculated as $\widehat{\theta}_c = \widehat{\theta} - \widehat{Bias}_b(\widehat{\theta})$.

We used a double-bootstrap procedure to calculate a 95% confidence interval for the bootstrap-bias corrected estimate of θ. Specifically, we bootstrapped from the data 300 times. In each bootstrap sample, we (double) bootstrapped 100 times and calculated the bootstrap bias-corrected estimate of θ. Percentiles of the bootstrap distribution of bias-corrected estimates were used to form the confidence interval.

References

Assareh, A., Moradi, M., and Volkert, L. (2008). A hybrid random subspace classifier fusion approach for protein mass spectra classification. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* pages 1–11.

Cole, S. and Hernán, M. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168,** 656–664.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, volume 57. CRC press.

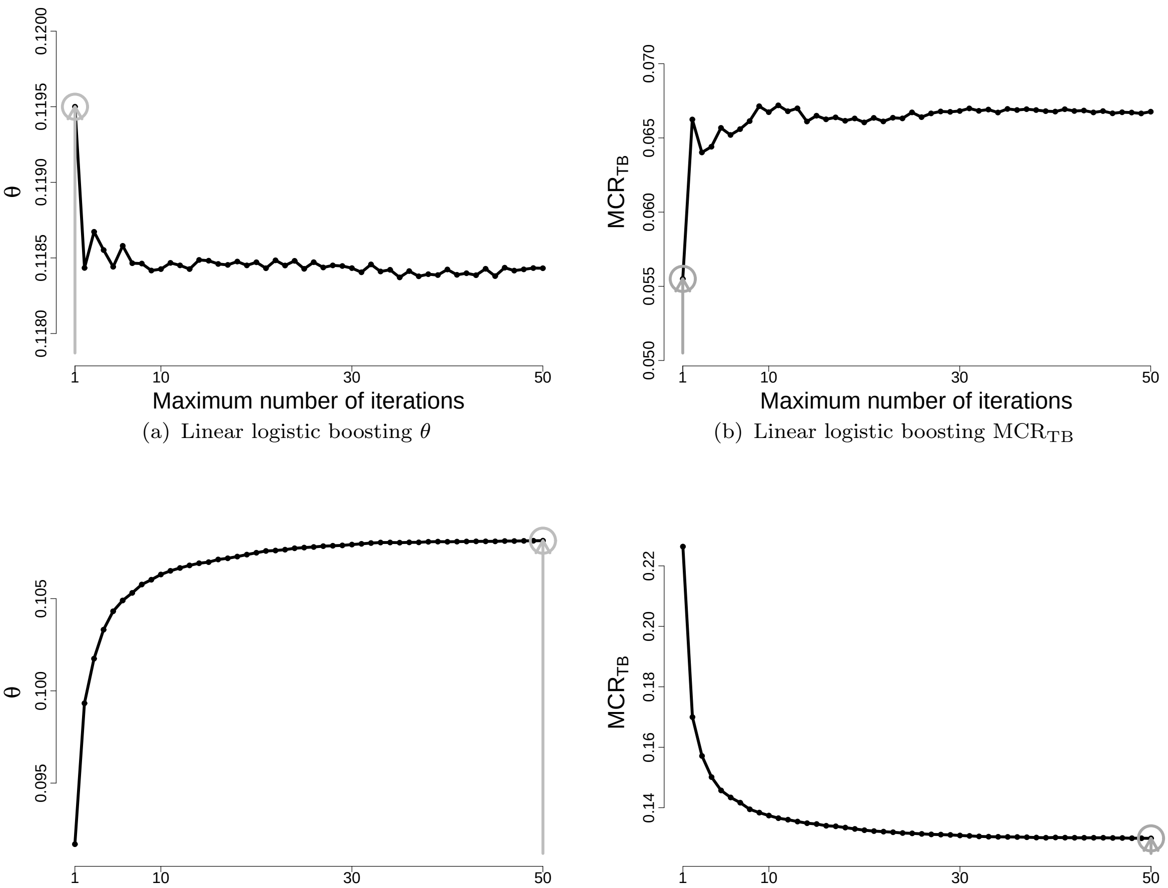
Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics* **28,** 337–407.

Lee, B., Lessler, J., and Stuart, E. (2011). Weight trimming and propensity score weighting. *Plos One* **6,** e18174.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* **11,** 169–198.

Potter, F. (1993). The effect of weight trimming on nonlinear survey estimates. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pages 758–763.

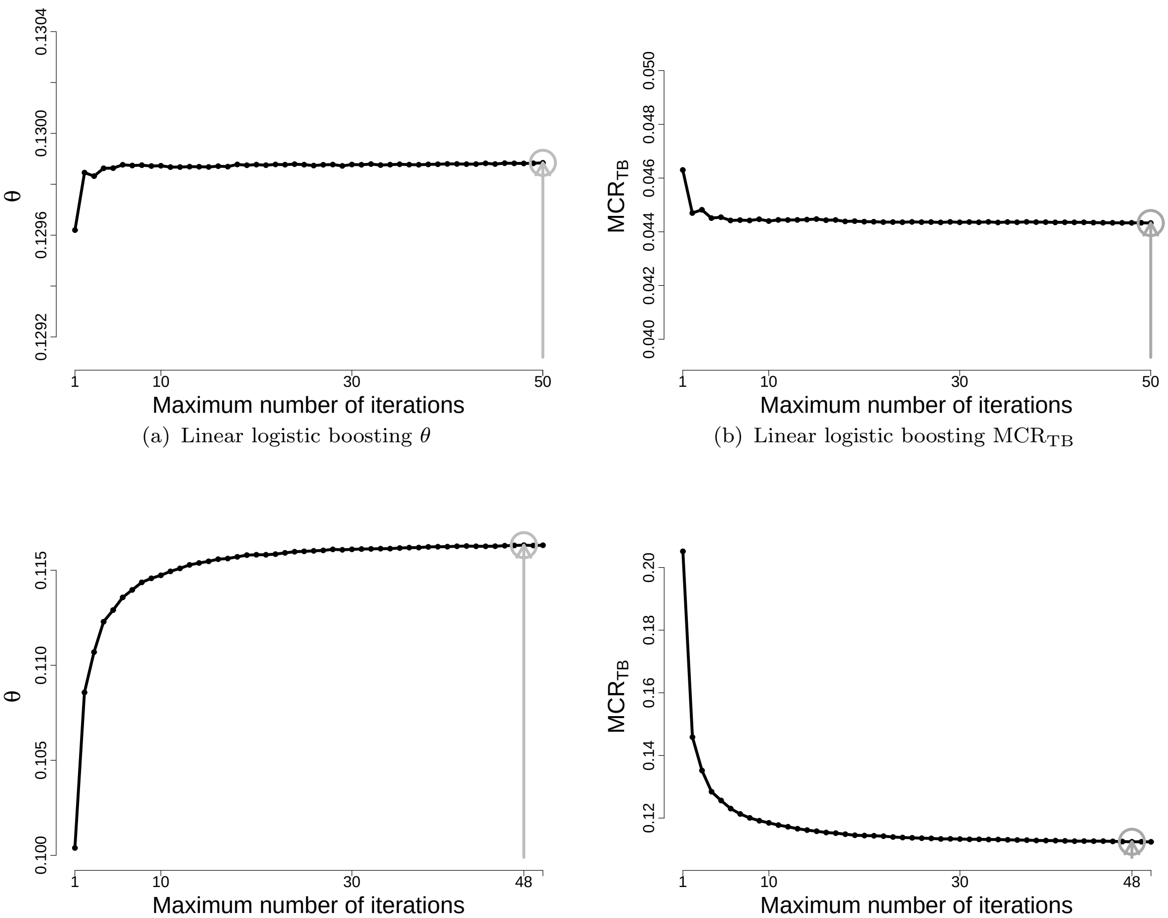
Received October 0000. *Revised February* 0000. *Accepted March* 0000.



(a) Linear logistic boosting θ (b) Linear logistic boosting $\mathrm{MCR}_{\mathrm{TB}}$

(c) Classification tree boosting θ (d) Classification tree boosting $\mathrm{MCR}_{\mathrm{TB}}$

Figure 1. Scenario 1 simulation results for the boosting method using different maximum number of iterations, $\mathrm{M}_{\mathrm{max}}$. Performance of marker combinations obtained using the following methods are compared: the boosting method described in Section 2.3 with linear logistic working model and the boosting method with classification tree working model. Mean θ and mean misclassification rate for treatment benefit ($\mathrm{MCR}_{\mathrm{TB}}$) in a large independent test data set over 1000 training data sets ($n = 500$) are shown for $\mathrm{M}_{\mathrm{max}} = 1, \ldots, 50$. The $\mathrm{M}_{\mathrm{max}} \le 50$ achieving the highest θ is indicated (grey arrow). The pre-specified convergence criterion for the logistic regression working model is $\|\widetilde{\beta}^{(k)} - \widetilde{\beta}^{(k-1)}\| \le 10^{-7}$, where $\widetilde{\beta}^{(k)}$ is the vector of estimated regression coefficients at the k^{th} iteration, or reaching $\mathrm{M}_{\mathrm{max}}$. For the non-parametric classification tree working model, the criterion is reaching $\mathrm{M}_{\mathrm{max}}$.



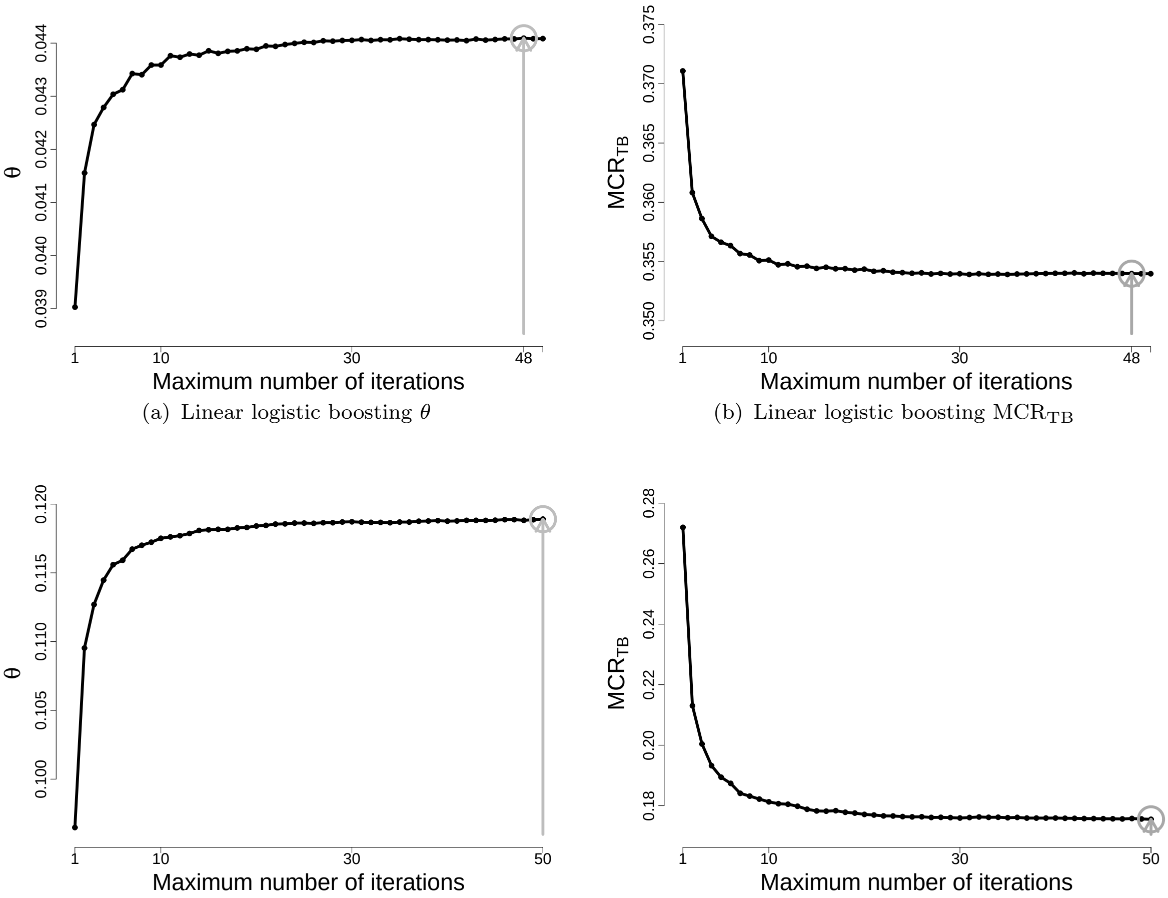
(a) Linear logistic boosting θ

(b) Linear logistic boosting $\mathrm{MCR}_{\mathrm{TB}}$

(c) Classification tree boosting θ

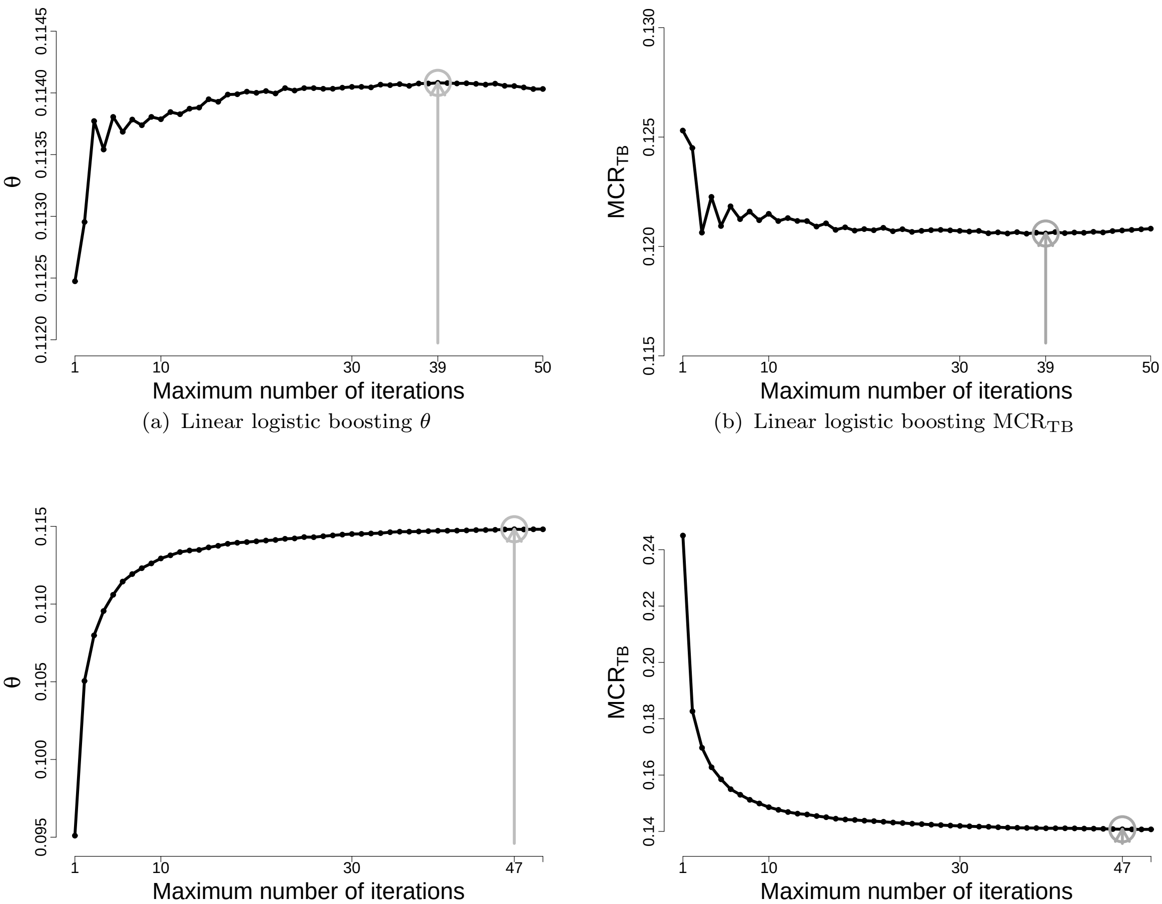
(d) Classification tree boosting $\mathrm{MCR}_{\mathrm{TB}}$

Figure 2. Scenario 3 simulation results for the boosting method using different maximum number of iterations, $\mathrm{M}_{\max}$. Performance of marker combinations obtained using the following methods are compared: the boosting method described in Section 2.3 with linear logistic working model and the boosting method with classification tree working model. Mean θ and mean misclassification rate for treatment benefit ($\mathrm{MCR}_{\mathrm{TB}}$) in a large independent test data set over 1000 training data sets ($n = 500$) are shown for $\mathrm{M}_{\max} = 1, \ldots, 50$. The $\mathrm{M}_{\max} \leq 50$ achieving the highest θ is indicated (grey arrow). The pre-specified convergence criterion for the logistic regression working model is $\|\widetilde{\beta}^{(k)} - \widetilde{\beta}^{(k-1)}\| \leq 10^{-7}$, where $\widetilde{\beta}^{(k)}$ is the vector of estimated regression coefficients at the k^{th} iteration, or reaching $\mathrm{M}_{\max}$. For the non-parametric classification tree working model, the criterion is reaching $\mathrm{M}_{\max}$.



(c) Classification tree boosting θ (d) Classification tree boosting $\mathrm{MCR_{TB}}$

Figure 3. Scenario 6 simulation results for the boosting method using different maximum number of iterations, $\mathrm{M_{max}}$. Performance of marker combinations obtained using the following methods are compared: the boosting method described in Section 2.3 with linear logistic working model and the boosting method with classification tree working model. Mean θ and mean misclassification rate for treatment benefit ($\mathrm{MCR_{TB}}$) in a large independent test data set over 1000 training data sets ($n = 500$) are shown for $\mathrm{M_{max}} = 1, \ldots, 50$. The $\mathrm{M_{max}} \leq 50$ achieving the highest θ is indicated (grey arrow). The pre-specified convergence criterion for the logistic regression working model is $\|\widetilde{\beta}^{(k)} - \widetilde{\beta}^{(k-1)}\| \leq 10^{-7}$, where $\widetilde{\beta}^{(k)}$ is the vector of estimated regression coefficients at the k^{th} iteration, or reaching $\mathrm{M_{max}}$. For the non-parametric classification tree working model, the criterion is reaching $\mathrm{M_{max}}$.



(a) Linear logistic boosting θ

(b) Linear logistic boosting $\mathrm{MCR}_{\mathrm{TB}}$

(c) Classification tree boosting θ

(d) Classification tree boosting $\mathrm{MCR}_{\mathrm{TB}}$

Figure 4. Scenario 7 simulation results for the boosting method using different maximum number of iterations, $\mathrm{M}_{\max}$. Performance of marker combinations obtained using the following methods are compared: the boosting method described in Section 2.3 with linear logistic working model and the boosting method with classification tree working model. Mean θ and mean misclassification rate for treatment benefit ($\mathrm{MCR}_{\mathrm{TB}}$) in a large independent test data set over 1000 training data sets ($n = 500$) are shown for $\mathrm{M}_{\max} = 1, \ldots, 50$. The $\mathrm{M}_{\max} \leq 50$ achieving the highest θ is indicated (grey arrow). The pre-specified convergence criterion for the logistic regression working model is $\|\widetilde{\beta}^{(k)} - \widetilde{\beta}^{(k-1)}\| \leq 10^{-7}$, where $\widetilde{\beta}^{(k)}$ is the vector of estimated regression coefficients at the k^{th} iteration, or reaching $\mathrm{M}_{\max}$. For the non-parametric classification tree working model, the criterion is reaching $\mathrm{M}_{\max}$.

Table 1

Simulation results for the boosting method using different weight functions, $\widetilde{w}\{\Delta(Y)\}$, and different maximum iteration parameters, M_{max}, with $n = 500$ observations. Simulation scenarios 1, 3, 6, and 7 are shown. Performance of marker combinations obtained using the following methods are compared: the boosting method described in Section 2.3 with linear logistic working model and the boosting method with classification tree working model. Mean and Monte Carlo standard deviation (SD) of θ are shown, along with the mean misclassification rate for treatment benefit (MCR_{TB}). The $M_{max} \leq 50$ achieving the highest mean θ (M_{Best}) is reported for the weight function $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{-\frac{1}{3}}$.

Scenario			Linear logistic boosting					Classification tree boosting				
		Weight $\widetilde{w}$	$\lvert\Delta(Y)\rvert^{-\frac{1}{3}}$		$\lvert\Delta(Y)\rvert^{-\frac{1}{10}}$	$e^{-\lvert\Delta(Y)\rvert}W_{\mathrm{A}}(Y)$	$e^{-\lvert\Delta(Y)\rvert}$	$\lvert\Delta(Y)\rvert^{-\frac{1}{3}}$		$\lvert\Delta(Y)\rvert^{-\frac{1}{10}}$	$e^{-\lvert\Delta(Y)\rvert}W_{\mathrm{A}}(Y)$	$e^{-\lvert\Delta(Y)\rvert}$
		maximum # of iterations	500	M_{Best}	500	500	500	500	M_{Best}	500	500	500
1	Optimal M=$\mathrm{M_{Best}}$			1					50			
	θ	Mean	0.1195	0.1195	0.1199	0.1199	0.1198	0.1083	0.1081	0.1009	0.0968	0.1023
		SD	0.0026	0.0026	0.0022	0.0023	0.0023	0.0065	0.0066	0.0107	0.0115	0.0098
	$\mathrm{MCR_{TB}}$	Mean	0.0555	0.0555	0.0521	0.0530	0.0524	0.1294	0.1299	0.1815	0.1701	0.1682
3	Optimal M=$\mathrm{M_{Best}}$			50					48			
	θ	Mean	0.1299	0.1299	0.1301	0.1305	0.1301	0.1162	0.1163	0.1095	0.0929	0.0977
		SD	0.0022	0.0022	0.0020	0.0018	0.0021	0.0065	0.0062	0.0095	0.0393	0.0378
	$\mathrm{MCR_{TB}}$	Mean	0.0444	0.0443	0.0420	0.0383	0.0423	0.1124	0.1124	0.1596	0.1731	0.1618
6	Optimal M=$\mathrm{M_{Best}}$			48					50			
	θ	Mean	0.0438	0.0441	0.0310	0.0234	0.0310	0.1186	0.1189	0.1064	0.1060	0.1040
		SD	0.0128	0.0122	0.0172	0.0191	0.0172	0.0106	0.0101	0.0167	0.0232	0.0101
	$\mathrm{MCR_{TB}}$	Mean	0.3542	0.3540	0.3739	0.3855	0.3739	0.1762	0.1755	0.2242	0.2119	0.2179
7	Optimal M=$\mathrm{M_{Best}}$			39					47			
	θ	Mean	0.1140	0.1141	0.1002	0.0963	0.1066	0.1151	0.1148	0.1089	0.1049	0.0962
		SD	0.0118	0.0117	0.0201	0.0241	0.0185	0.0094	0.0094	0.0137	0.0139	0.0347
	$\mathrm{MCR_{TB}}$	Mean	0.1207	0.1206	0.1752	0.1879	0.1506	0.1394	0.1408	0.1851	0.1830	0.2010

$W_{\mathrm{A}}(Y) = \exp\left\{-\frac{1}{2}\log\left(\frac{1-err}{err}\right) \times (2D-1)(2\widehat{D}-1)\right\}$, where

D denotes the binary outcome (0 or 1), $\widehat{D} = \mathbf{1}\{\widehat{P}(D=1|T,Y) > 0.5\}$ denotes the predicted outcome in the previous stage, and $err = P(D \neq \widehat{D})$.

Table 2

The best weight function and the maximum number of iterations for the boosting method in the breast cancer data. Models including the modified risk score (MRS); genes G_1, G_2 and G_3; and genes G_4, G_5 and $G_4 \times G_5$ are shown. Weight functions of the form $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^d$ were considered. The best weight function and the maximum number of iterations are determined based on the average θ over 10 replications of 5-fold cross-validation.

Marker set (Y) / Working model	Linear logistic		Classification tree with interactions	
	d in $\widetilde{w}(\Delta(Y)) = \|\Delta(Y)\|^d$	Maximum # of iterations (M_{max})	d in $\widetilde{w}(\Delta(Y)) = \|\Delta(Y)\|^d$	Maximum # of iterations (M_{max})
MRS	-1.83	100	-0.82	15
(G_1, G_2, G_3)	-0.33	270	-0.14	20
$(G_4, G_5, G_4 \times G_5)$	-1.85	150	-1.85	250

Table 3

Simulation results for the boosting method using different choices for the maximum weight, C_M. Simulation scenarios 1, 3, 6, and 7 with 500 observations are examined. The boosting method described in Section 2.3 is applied with linear logistic working model, $\widetilde{w}\{\Delta(Y)\} = |\Delta(Y)|^{-\frac{1}{3}}$, and $M_{max} = 500$. Mean θ and mean misclassification rate for treatment benefit (MCR_{TB}) in a large independent test data set across 1000 training data sets are shown.

		Maximum weight (C_M)		
		300	500	1000
Scenario 1	θ	0.11949	0.11949	0.11949
	MCR_{TB}	0.05557	0.05552	0.05557
Scenario 3	θ	0.12988	0.12988	0.12988
	MCR_{TB}	0.04437	0.04436	0.04436
Scenario 6	θ	0.04383	0.04384	0.04384
	MCR_{TB}	0.35423	0.35418	0.35419
Scenario 7	θ	0.11408	0.11405	0.11408
	MCR_{TB}	0.12060	0.12071	0.12059