Web-based Supplementary Materials for Empirical Bayes Estimation and Prediction Using Summary-Level Information From External Big Data Sources Adjusting for Violations of Transportability

Jason P. Estes, Bhramar Mukherjee, Jeremy M. G. Taylor
University of Michigan, MI 48109, USA

Table 1: Estimated bias, standard deviation (SD) and mean squared error (MSE) of parameter estimates in linear regression settings I, II, III and IV specified in Table 1 of the main text with full model (13) and reduced model (14) based on 1000 simulation runs. LR denotes the maximum likelihood estimates of (13) fitted to the internal data, EB denotes our empirical Bayes estimator defined in (4) and CML denotes the constrained maximum likelihood estimator proposed in Chatterjee at al. (2016).

		BIAS				SD				MSE			
Setting	Method	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
I	LR	-.001	-.006	.008	-.001	.205	.200	.198	.187	.042	.040	.039	.035
	EB	-.009	-.002	.006	-.001	.166	.167	.198	.187	.028	.028	.039	.035
	CML	-.029	.005	.012	.003	.118	.137	.198	.187	.015	.019	.039	.035
II	LR	.004	-.019	.020	.001	.202	.194	.201	.190	.041	.038	.041	.036
	EB	.081	.045	-.028	-.042	.198	.175	.204	.194	.046	.033	.042	.039
	CML	.609	.106	.081	.052	.128	.147	.202	.190	.387	.033	.047	.039
III	LR	.002	.000	.014	.004	.280	.268	.194	.161	.078	.072	.038	.026
	EB	-.018	.069	.008	-.006	.264	.264	.194	.163	.070	.075	.038	.026
	CML	-.106	.272	.025	.020	.236	.231	.195	.162	.067	.127	.038	.026
IV	LR	.003	-.004	-.003	-.005	.279	.263	.194	.179	.078	.069	.038	.032
	EB	.017	.062	-.037	-.077	.277	.275	.200	.207	.077	.080	.041	.049
	CML	.215	.521	.041	.058	.227	.232	.195	.205	.098	.325	.040	.045

Table 2: Estimated bias, standard deviation (SD) and mean squared error (MSE) of parameter estimates in logistic regression settings I, II, III and IV specified in Table 1 of the main text with full model (13) and reduced model (14) based on 1000 simulation runs. LR denotes the maximum likelihood estimates of (13) fitted to the internal data, EB denotes our empirical Bayes estimator defined in (4) and CML denotes the constrained maximum likelihood estimator proposed in Chatterjee at al. (2016).

		BIAS				SD				MSE			
Setting	Method	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
I	LR	-.003	-.003	.005	.003	.083	.081	.078	.070	.007	.007	.006	.005
	EB	.002	-.003	.005	.003	.067	.066	.078	.070	.005	.004	.006	.005
	CML	.014	-.005	.005	.004	.037	.039	.078	.070	.002	.002	.006	.005
II	LR	-.004	-.004	.003	-.002	.080	.081	.081	.071	.006	.007	.007	.005
	EB	.015	.005	.002	-.004	.081	.077	.081	.071	.007	.006	.007	.005
	CML	.229	.102	.004	.002	.039	.042	.081	.071	.054	.012	.007	.005
III	LR	-.005	-.005	.003	.001	.105	.115	.081	.043	.011	.013	.006	.002
	EB	-.008	.011	.001	-.008	.105	.117	.081	.044	.011	.014	.006	.002
	CML	-.068	.216	.004	.009	.078	.094	.081	.043	.011	.056	.007	.002
IV	LR	-.004	-.002	.001	.001	.103	.111	.083	.049	.011	.012	.007	.002
	EB	.011	.015	-.001	-.012	.102	.108	.083	.064	.011	.012	.007	.004
	CML	.120	.211	.005	.011	.070	.069	.083	.061	.019	.049	.007	.004

Table 3: Characteristics of study participants in the external, internal and validation data sets of our data application in Section 4 of the main text.

	Thompson (External)		Training (Internal)		Validation	
Age	N (5519)	(%)	N (711)	(%)	N (1225)	(%)
< 55	0	0.0	119	16.7	195	15.9
[55,60)	38	0.7	137	19.3	195	15.9
[60, 64)	1143	20.7	132	18.6	188	15.3
[65, 69)	1741	31.5	117	16.5	237	19.3
≥ 70	2597	47.1	142	20	333	27.2
NA			1	0.1		
Family History of PCa						
No	4599	83.3	546	76.8	940	76.7
Yes	920	16.7	162	22.8	230	18.8
NA	–	–	3	0.4	55	4.5
Race						
White	5276	95.6	568	79.9	512	41.8
African American	175	3.2	73	10.3	82	6.7
Other/Unknown	68	1.2	70	9.8	631	51.5
Number of previous negative biopsies						
0	4873	88.3	196	27.6	246	20.1
$>= 1$	753	13.6	515	72.4	977	79.8
NA	–	–	–	–	2	0.2
PSA Level (ng/mL)						
0 - 1	1963	35.6	17	2.4	54	4.4
1.1 - 2	1640	29.7	30	4.2	97	7.9
2.1 - 3	775	14.0	53	7.5	126	10.3
3.1 - 4	510	9.2	96	13.5	170	13.9
4.1 - 6	481	8.7	274	38.5	419	34.2
> 6	150	2.7	241	33.9	359	29.3
HG PCa	257	4.7	192	27.0	224	18.3

Table 4: Parameter estimates of model (17) fitted to our data application using our empirical Bayes (EB) estimator defined in (4) of the main text and the constrained maximum likelihood (CML) defined in Chatterjee et al. (2016). LR^{F} and LR^{R} denotes the parameter estimates of the full (17) and reduced models (16) defined in the main text respectively fitted to the internal data set or the validation data set of our data application.

	Prediction Model		Internal Data		Validation Data		External Data
Parameter	EB	CML	LR^{F}	LR^{R}	LR^{F}	LR^{R}	PCPTrc
Intercept	-6.076	-7.505	-6.097	-4.740	-7.269	-6.542	-6.246
lpsa	0.860	1.188	0.867	0.958	0.924	0.934	1.293
age	0.010	0.007	0.010	0.034	0.032	0.054	0.031
dre	0.975	0.756	0.984	1.139	0.607	0.623	1.001
priobiop	-1.024	-0.220	-1.051	-1.154	-0.942	-1.054	-0.363
aa	0.184	0.699	0.197	0.460	0.050	0.116	0.960
lt2erg	0.528	0.518	0.524	–	0.318	–	–
lpca3	0.129	0.127	0.128	–	0.163	–	–

Table 5: Simulation results. Proportion of 1,000 simulation runs for which absolute estimation error in estimation for method A (CML or EB) was less than absolute estimation error in estimation for method B (LR or CML) when a covariate vector is randomly drawn from the external covariate distribution (Ext) or internal covariate distribution (Int) in each regression settings I, II, III and IV defined in Table 1 of the main text.

Full Model (12)

	Standard Linear Regression							Standard Logistic Regression						
Setting	I	II		III		IV		I	II		III		IV	
Distribution	Ext	Ext	Int	Ext	Int	Ext	Int	Ext	Ext	Int	Ext	Int	Ext	Int
CML/LR	.65	.08	.08	.64	.65	.22	.24	.69	.10	.09	.72	.72	.26	.31
EB/LR	.71	.42	.45	.70	.70	.47	.48	.75	.48	.48	.79	.77	.54	.57
EB/CML	.42	.93	.93	.41	.41	.79	.79	.40	.91	.92	.34	.36	.76	.72

Full Model (13)

	Standard Linear Regression							Standard Logistic Regression						
Setting	I	II		III		IV		I	II		III		IV	
Distribution	Ext	Ext	Int	Ext	Int	Ext	Int	Ext	Ext	Int	Ext	Int	Ext	Int
CML/LR	.64	.15	.14	.46	.42	.31	.32	.69	.16	.15	.30	.34	.32	.30
EB/LR	.68	.47	.48	.57	.52	.45	.44	.76	.52	.50	.50	.50	.49	.52
EB/CML	.42	.87	.88	.58	.62	.71	.68	.36	.85	.86	.70	.68	.70	.72

Table 6: Monte Carlo (MC) and Bootstrap (BS) standard deviation (SD) estimates in the linear and logistic regression setting I with full model (12) and reduced model (14) averaged over 1000 simulation runs. In each Monte Carlo run, 500 bootstrap samples were used.

		Bootstrap SD				MC
Method	Statistic	Q1	Q2	Q3	Mean	SD
Linear	$\widehat{\beta}_0$	0.155	0.163	0.174	0.165	0.153
	$\widehat{\beta}_1$	0.166	0.174	0.184	0.176	0.169
	$\widehat{\beta}_2$	0.191	0.198	0.206	0.199	0.199
Logistic	$\widehat{\beta}_0$	0.060	0.064	0.068	0.065	0.060
	$\widehat{\beta}_1$	0.066	0.070	0.075	0.070	0.065
	$\widehat{\beta}_2$	0.078	0.080	0.083	0.080	0.079

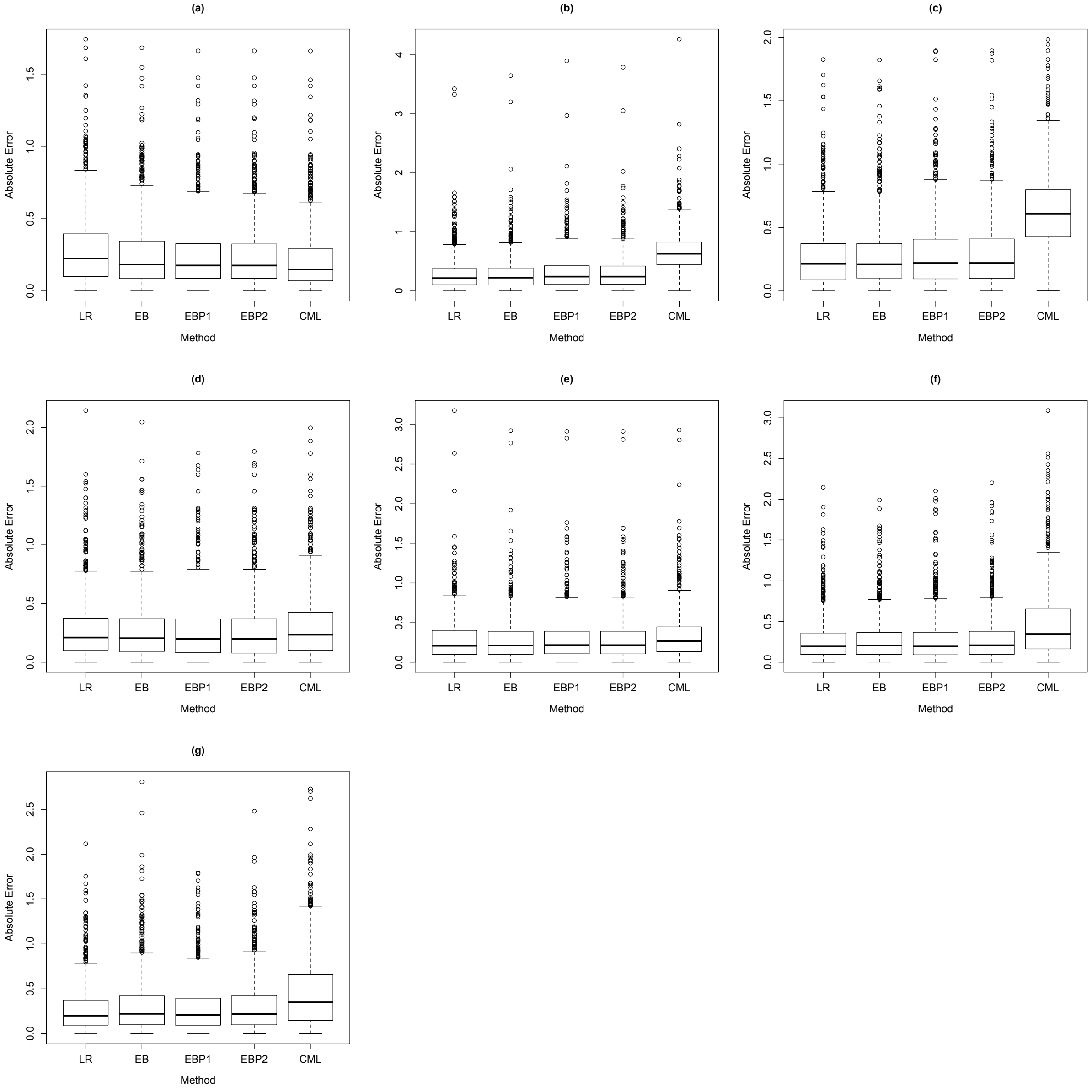

Figure 1: Box plots of absolute estimation error defined by (a),(b),(d),(f) $|\widehat{M}_{E,r} - W_{E,r}^{\mathrm{T}}\beta|$ and (a),(c),(e),(g) $|\widehat{M}_{I,r} - W_{I,r}^{\mathrm{T}}\beta|$ based on $r = 1, \ldots, 1000$ simulation runs in the standard linear regression settings I (a), II (b and c), III (d and e), IV (f and g) specified in Table 1 of the main text with full model (13) and reduced model (14) $W_{E,r}^{\mathrm{T}}$ is a covariate vector drawn from the external population, $W_{I,r}^{\mathrm{T}}$ is drawn from the internal population, $\widehat{M}_{E,r}$ and $\widehat{M}_{I,r}$ are estimates of the conditional mean of Y given (X, Z) in the external and internal populations respectively resulting from maximum likelihood (LR), our empirical Bayes estimators EB, EBP1, and EBP2 or the constrained maximum likelihood estimator CML.

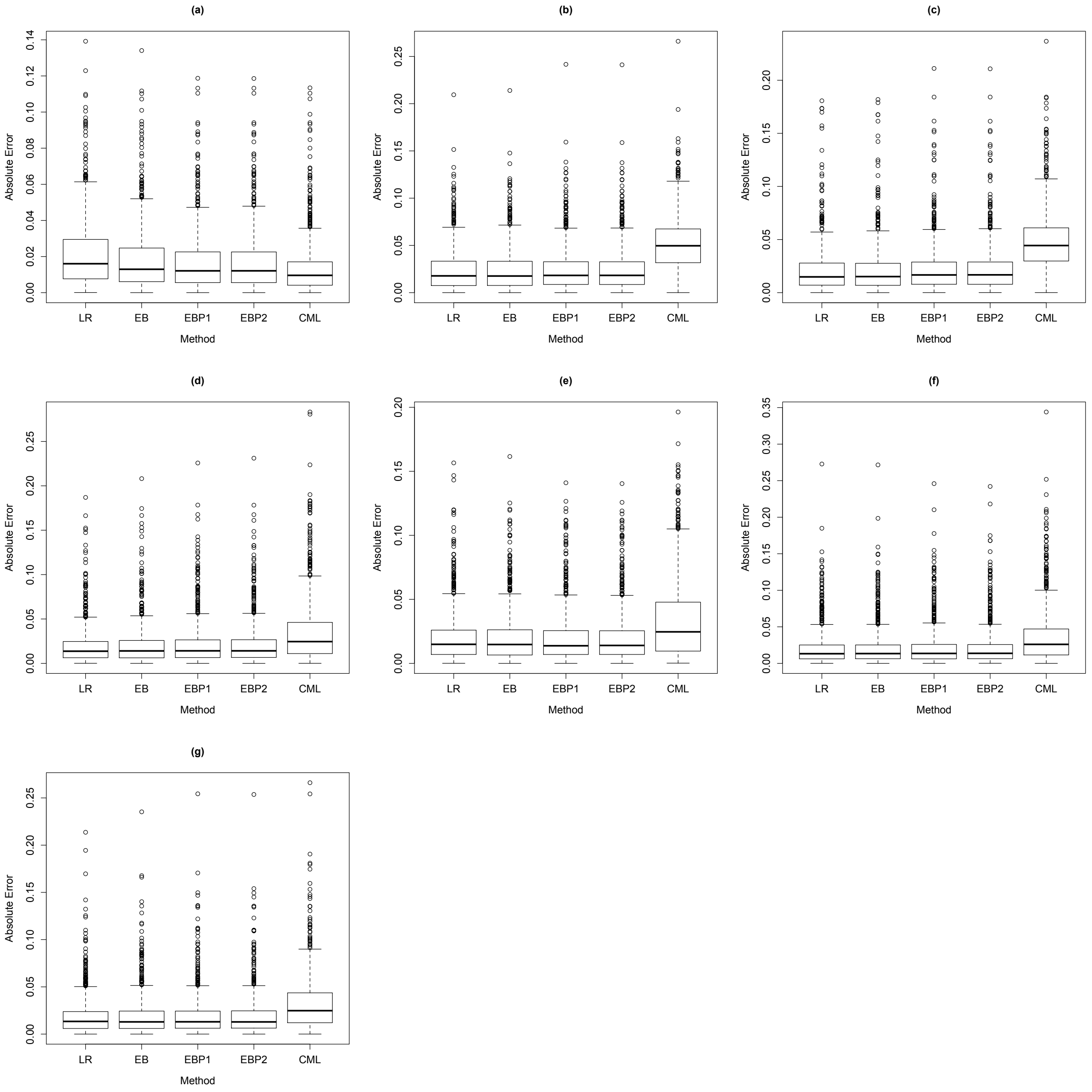

Figure 2: Box plots of absolute estimation error defined by (a),(b),(d),(f) $|\widehat{M}_{E,r} - g^{-1}(W_{E,r}^{\mathrm{T}}\beta)|$ and (a),(c),(e),(g) $|\widehat{M}_{I,r} - g^{-1}(W_{I,r}^{\mathrm{T}}\beta)|$ based on $r = 1, \ldots, 1000$ simulation runs in the standard logistic regression settings I (a), II (b and c), III (d and e), IV (f and g) specified in Table 1 of the main text with full model (13) and reduced model (14) $W_{E,r}^{\mathrm{T}}$ is a covariate vector drawn from the external population, $W_{I,r}^{\mathrm{T}}$ is drawn from the internal population, $\widehat{M}_{E,r}$ and $\widehat{M}_{I,r}$ are estimates of the conditional mean of Y given (X, Z) in the external and internal populations respectively resulting from maximum likelihood (LR), our empirical Bayes estimators EB, EBP1, and EBP2 or the constrained maximum likelihood estimator CML.

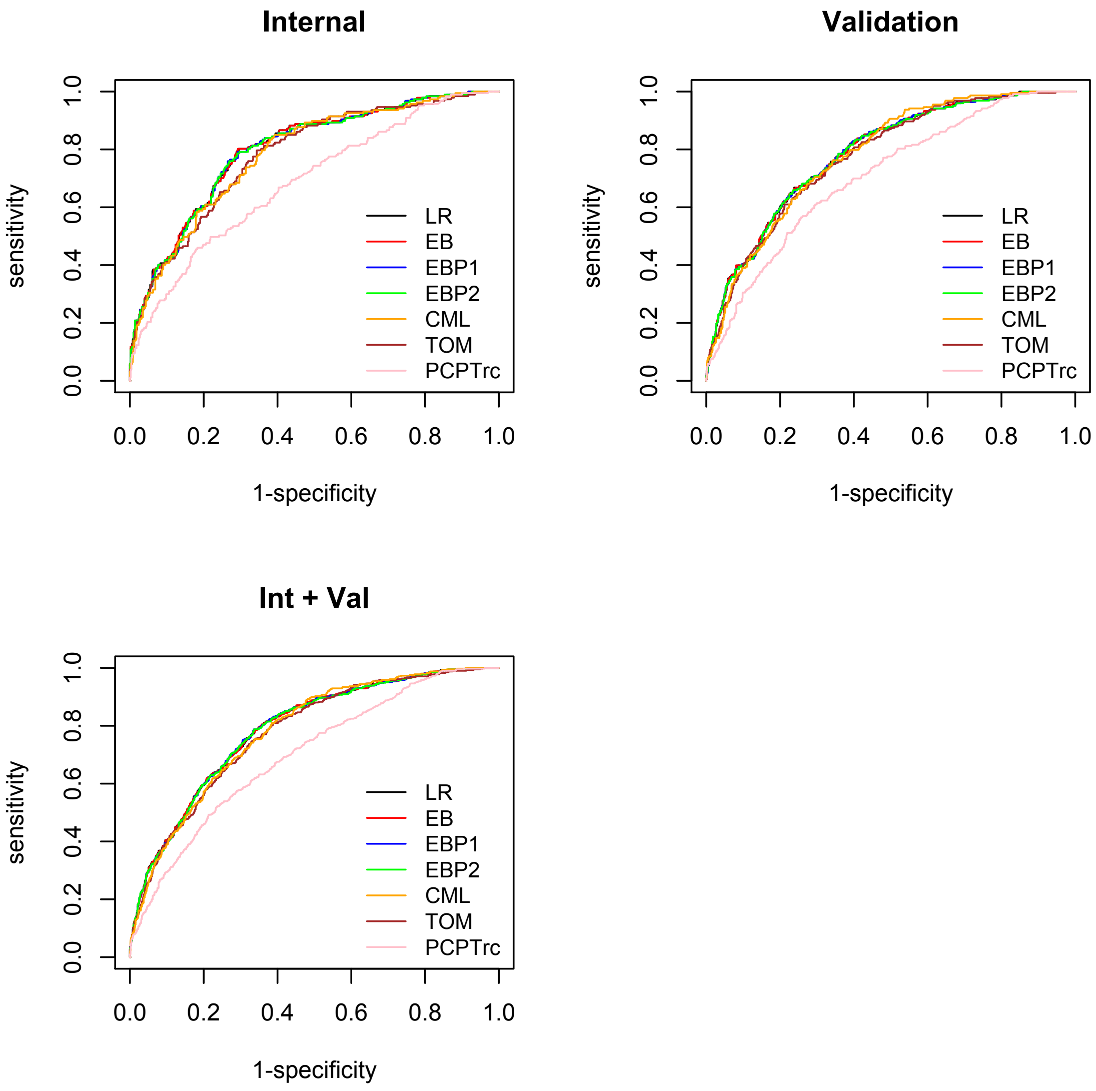

Figure 3: Receiver-operator curves resulting from maximum likelihood, our empirical Bayes estimators EB, EBP1 and EBP1 defined in Section 2.3, the constrained maximum likelihood estimator CML proposed in Chatterjee et al. (2016), the model (TOM) proposed in Tomlins et al. (2016), or the PCPTrc 1.0 calculator applied to the internal, validation and combined data sets of our data application.

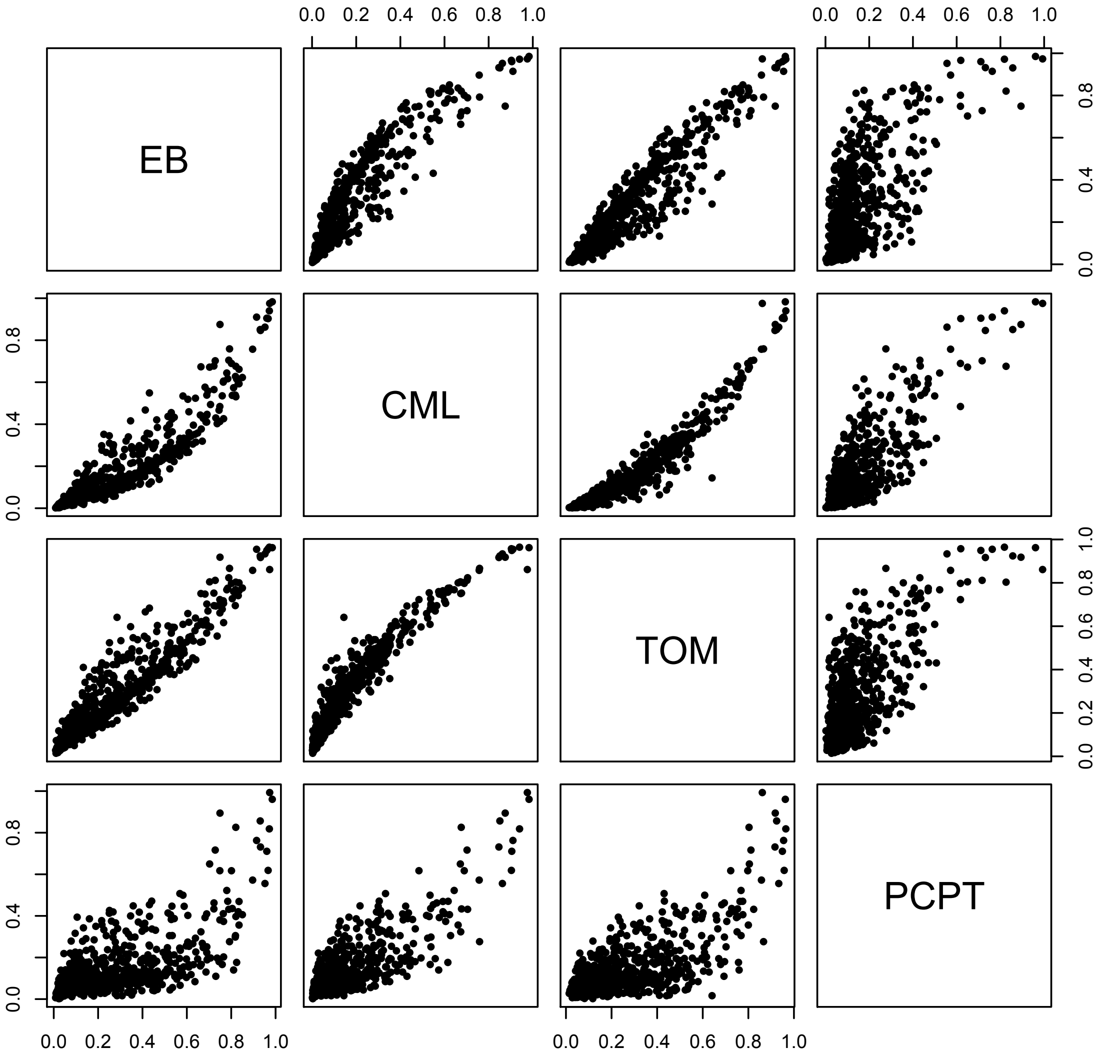

Figure 4: Scatterplot matrix of predicted outcomes resulting from our empirical Bayes estimator EB defined in equation (4), the constrained maximum likelihood estimator CML proposed in Chatterjee et al. (2016), the model (TOM) proposed in Tomlins et al. (2016), or the PCPTrc 1.0 calculator applied to the internal data application.

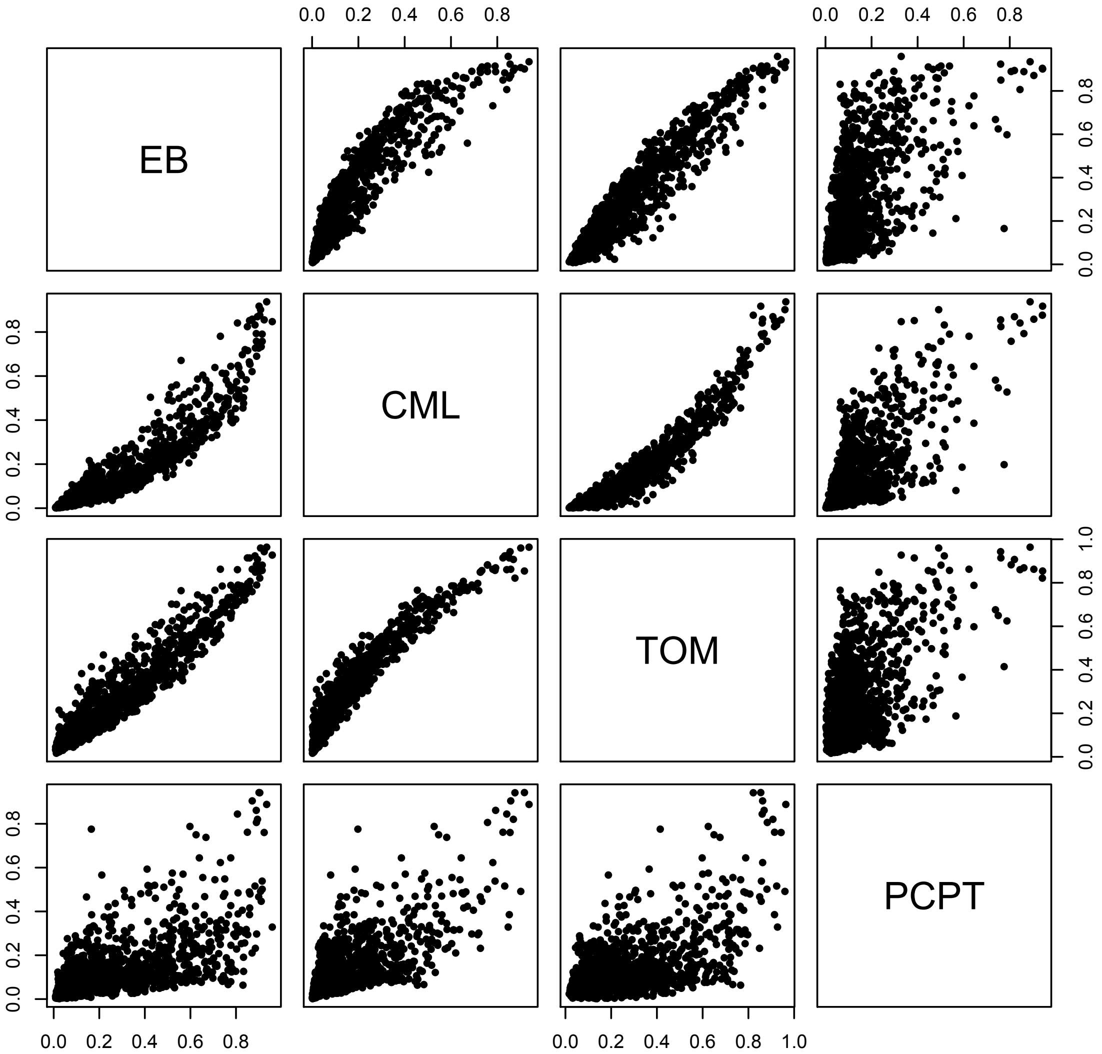

Figure 5: Scatterplot matrix of predicted outcomes resulting from our empirical Bayes estimator EB defined in equation (4), the constrained maximum likelihood estimator CML proposed in Chatterjee et al. (2016), the model (TOM) proposed in Tomlins et al. (2016), or the PCPTrc 1.0 calculator applied to the validation data application.