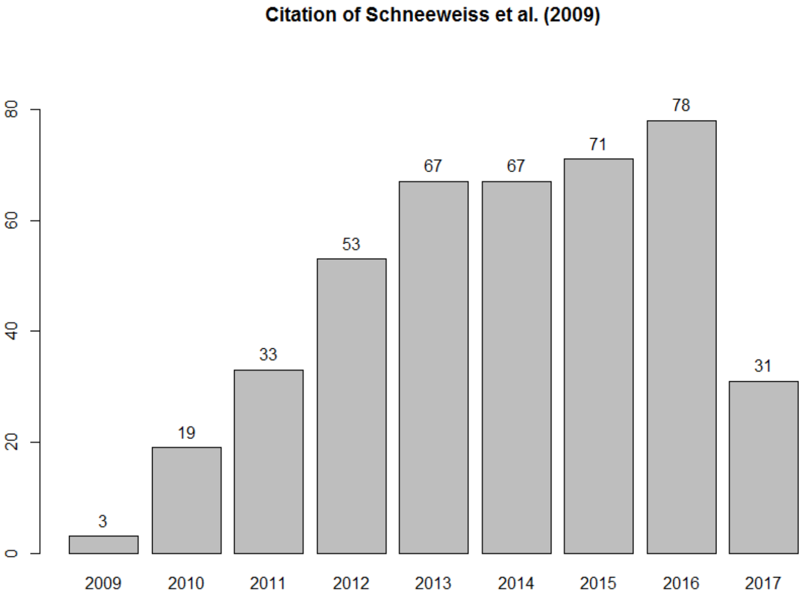
A eAPPENDIX

Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; OR, odds ratio; RD, risk difference; EC, empirical covariate.

A.1 Popularity of High-dimensional propensity score adjustment

Schneeweiss and his colleagues[1] argued that adjusting for additional proxy information from the health administrative dataset via a PS model should further reduce bias in estimating the treatment effects. Considering these proxy data in the analysis, they showed data analysis examples where hdPS analysis results were closer to randomized controlled trial results compared to the conventional PS analysis results.



eFigure A.1: Citation of the Schneeweiss, S., et al. paper (published in Epidemiology, 2009 that originally outlined the High-dimensional propensity score algorithm) over the years. Citation data collected from the Google scholar in 24th April, 2017.

A.2 Investigator-specified predefined covariates

Potential confounders identified as predefined covariates for the study are demographic characteristics (e.g. age, sex), time variables (e.g. year of cohort entry), clinical characteristics (e.g., smoking, alcohol use, obesity), comorbidities (e.g. diabetes mellitus, atrial fibrillation, coronary artery disease recorded > 30 days before the index MI, acute coronary syndrome, cerebrovascular disease, congestive heart failure, chronic obstructive pulmonary disease, hypertension, hypercholesterolemia, peripheral vascular disease, previous coronary revascularization, previous stroke, previous MI, recorded > 30 days before the index MI, and previous medications prescribed. Previous medications prescribed included aspirin, angiotensin-converting enzyme (ACE) inhibitors, angiotensin receptor blockers (ARBs), beta-blockers, calcium-channel blockers, diuretics, fibrates, non-steroidal anti-inflammatory drugs (NSAIDs). We also constructed variables for the number of prescriptions issued and the number of hospitalizations in the previous year, which are two proxies for overall health. Age, the number of hospitalization, and prescription count were categorized into groups, and they were considered as dummy variables along with the year of cohort entry.

A.3 Baseline Characteristics of Post MI Patients with respect to Statins Use

On average, the statin user group is younger, more of them are male, more are smokers, more obese, and less are diabetic patients.

eTable A.1: Baseline Characteristics for important confounders

	No Statin	Statin
Cohort size	13671	19121
Age*(yrs, SD)	73.14 (13.87)	65.99 (13.22)
Male (%)	7783 (56.9)	13021 (68.1)
Smoking (%)	8580 (62.8)	13003 (68.0)
Obesity (%)	1620 (11.8)	3051 (16.0)
Comorbidities (%)		
Diabetes mellitus	1939 (14.2)	1849 (9.7)

* Age is considered as a continuous variable in the plasmode simulation and the data analysis.

eTable A.2: Baseline Characteristics for additional investigator-specified confounders

	No Statin	Statin
Alcohol use (%)	169 (1.2)	332 (1.7)
Year of entry (%)		
1998	905 (6.6)	389 (2.0)
1999	1304 (9.5)	698 (3.7)
2000	1409 (10.3)	970 (5.1)
2001	1483 (10.8)	1190 (6.2)
2002	1282 (9.4)	1559 (8.2)
2003	1158 (8.5)	1738 (9.1)
2004	967 (7.1)	1690 (8.8)
2005	797 (5.8)	1549 (8.1)
2006	737 (5.4)	1598 (8.4)
2007	708 (5.2)	1560 (8.2)
2008	657 (4.8)	1470 (7.7)
2009	687 (5.0)	1472 (7.7)
2010	697 (5.1)	1473 (7.7)
2011	718 (5.3)	1389 (7.3)
2012	162 (1.2)	376 (2.0)
Comorbidities (%)		
Atrial fibrillation	2418 (17.7)	1763 (9.2)
Coronary artery disease	2608 (19.1)	1489 (7.8)
Acute coronary syndrome	1344 (9.8)	2412 (12.6)
Cerebrovascular disease	1048 (7.7)	607 (3.2)
Congestive heart failure	3147 (23.0)	2580 (13.5)
Chronic obstructive pulmonary disease	1336 (9.8)	1233 (6.4)
Hypertension	4428 (32.4)	6554 (34.3)
Hypercholesterolemia	1473 (10.8)	4040 (21.1)
Peripheral vascular disease	610 (4.5)	511 (2.7)
Previous coronary revascularization	2076 (15.2)	6875 (36.0)
Previous stroke	690 (5.0)	341 (1.8)
Previous MI	891 (6.5)	380 (2.0)
Previous medications prescribed (%)		
Aspirin	6546 (47.9)	17127 (89.6)
Ace inhibitors	4518 (33.0)	14533 (76.0)
arBs	768 (5.6)	1269 (6.6)
Beta-blockers	4444 (32.5)	15228 (79.6)
calcium-channel blockers	3231 (23.6)	4303 (22.5)
Diuretics	5723 (41.9)	6076 (31.8)
Fibrates	177 (1.3)	125 (0.7)
nSaiDs	2794 (20.4)	4232 (22.1)
Prescription count*(SD)	8.67 (6.69)	9.99 (5.25)
# of hospitalization*(SD)	1.55 (2.02)	1.45 (0.89)

* Prescription count and number of hospitalization are considered as continuous variables in the plasmode simulation and the data analysis.

A.4 Creating empirical covariates

To deal with residual confounding, we utilized additional information from the same database as proxies for unmeasured confounding. According to the proposed algorithm[1], to convert them into appropriate covariates, we follow the following steps. Before treatment initiation in the dataset, a temporal window of 1-year is set when we collect the baseline proxy covariates. This window is known as the “Pre-treatment covariate assessment period”[1]. In this time-period, we receive proxy data columns from 4 data sources or dimensions: (a) general practice data (b) diagnosis data (c) procedure data (d) medication data. We only allow for the top 200 most prevalent codes. Schuster et al. (2015) showed that confounder variables with low prevalence may become influential when the prevalence of either exposure category is low[2]. Therefore, there is no theoretical justification to follow this ‘prevalence-targeted pre-selection’ step[2] in the hsPS algorithm. To show the detrimental impact on the estimated risk ratios from the hdPS approach, they used a hypothetical example of a point-exposure study with a binary outcome. However, to the best of our knowledge, there hasn’t been a systematic study yet with high-dimensional empirical cohorts that compared the impact of excluding this step from the hdPS algorithm. The authors did point out that in the large pharmacoepidemiological studies, the frequencies of exposed patients are generally sufficient in practice to allow researchers to reliably estimate the measure of effect using the hdPS or even the general PS method[2]. As this prevalence-targeted pre-selection step can be useful in reducing the already high dimensional problem in the dataset and thereby, making the data size manageable (before series of prioritization calculations are conducted), researchers continue to use this step heuristically in studies, except for those with infrequent exposures[3]. Each of these column data is classified into 3 levels of within-patient frequency of occurrence (i.e., once, sporadic and frequent) during the baseline period. Based on presence versus absence of the respective occurrence levels, binary proxy or empirical covariates are created.

A.5 Scores used for Prioritization

Let c be a binary empirical covariate, D be the binary indicator for outcome and E be the exposure status (also binary). The bias formula proposed by Bross (1966) is provided as follows:

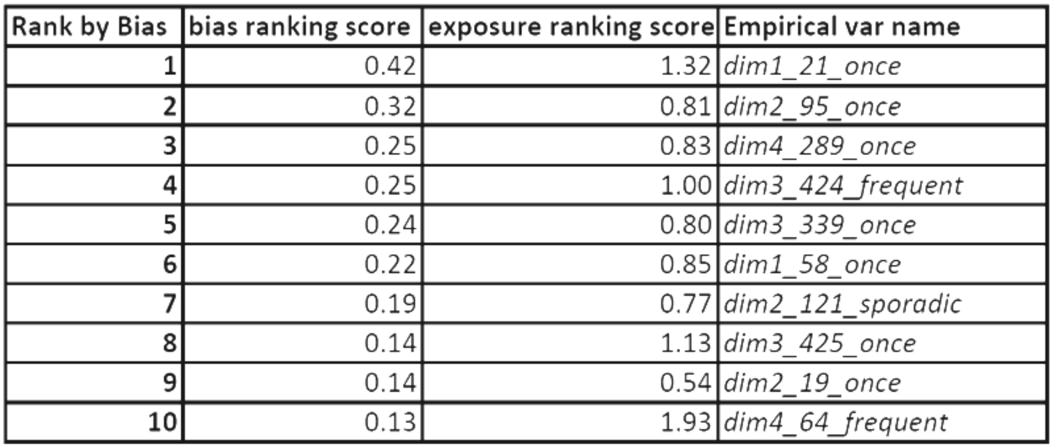
$$Bias_M = \begin{cases} \frac{P_{c_1}(RR_{CD}-1)+1}{P_{c_0}(RR_{CD}-1)+1}, & \text{if } RR_{CD} \geq 1 \\ \frac{P_{c_1}(\frac{1}{RR_{CD}}-1)+1}{P_{c_0}(\frac{1}{RR_{CD}}-1)+1}, & \text{otherwise} \end{cases} \tag{A.1}$$

where, P_{c_1} = prevalence among treated, P_{c_0} = prevalence among untreated, P_{cD_1} = prevalence among dead, P_{cD_0} = prevalence among alive. Here, $RR_{CD} = P_{cD_1}/P_{cD_0}$.

For the bias-based hdPS algorithm, $\log(Bias_M)$ is used as a rank score to determine priority (higher the score, more potential for confounding). The hdPS algorithm calculates the "bias score" ($Bias_M$) according to this bias formula proposed by Bross[4]. This formula is used to calculate the association between an empirical-covariate and the outcome, adjusting for the exposure prevalence imbalance. According to the magnitude of the absolute log-bias score, all the empirical-covariates are ranked. Such ranking is known as 'bias-based' ranking. For 'exposure-based' hdPS algorithm, the rank score is $\log(RR_{CE})$, where,

$$RR_{CE} = \frac{P_{c_1}}{P_{c_0}}. \tag{A.2}$$

eFigure A.2 shows top 10 empirical variables chosen by the bias-based ranking in a hypothetical hdPS analysis. Ranking in terms of exposure-based metric would result in differnt set of empirical variables.



Rank by Bias	bias ranking score	exposure ranking score	Empirical var name
1	0.42	1.32	*dim1_21_once*
2	0.32	0.81	*dim2_95_once*
3	0.25	0.83	*dim4_289_once*
4	0.25	1.00	*dim3_424_frequent*
5	0.24	0.80	*dim3_339_once*
6	0.22	0.85	*dim1_58_once*
7	0.19	0.77	*dim2_121_sporadic*
8	0.14	1.13	*dim3_425_once*
9	0.14	0.54	*dim2_19_once*
10	0.13	1.93	*dim4_64_frequent*

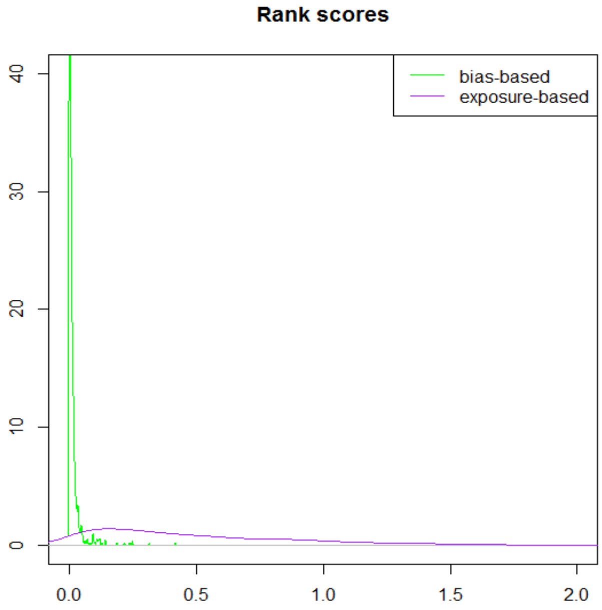
eFigure A.2: Ranking by log-bias score

As shown in eFigure A.3, the densities of rank scores are also generally different.

Note that, the investigator-specified variables do not go through selection process in the hdPS methods in the above mentioned prioritization process. Only the empirical covariates are prioritized and selected accordingly.

A.6 Software for the Machine learning algorithm

For fitting LASSO and elastic net, we used `cv.glmnet` function from the `glmnet` package in `R` varying `alpha` values (`alpha` = 1 for LASSO and `alpha` = 0.5 for our elastic net fitting) and



eFigure A.3: Density of rank scores

setting the following options `nfolds = 5` and `nlambda = 100`. For example: for a given binary outcome vector `y` and model matrix `x`, we can run the elastic net model as follows:

```
require(glmnet)
fit.k.fold <- cv.glmnet(x, y, family = "binomial", alpha = 0.5,
                        standardize = TRUE, lambda = NULL,
                        type.measure = 'deviance', nfolds = 5,
                        nlambda = 100)
pred <- predict(fit.k.fold$glmnet.fit, newx = x, type = 'response',
                  s = fit.k.fold$lambda.min)
fit <- list(object = fit.k.fold, useMin = TRUE)
fit$pred <- pred
fit$varname <- dimnames(coef(fit.k.fold))[[1]]
```

For the above elastic net model fitting, it is possible to choose an optimum `alpha` value by cross-validating over a grid of candidate values. But for sake of reducing computational burden, we chose to use a fixed `alpha` = 0.5 value. Franklin et al. (2015) is a very useful reference for fitting LASSO (i.e., `alpha` = 1 in `glmnet`; see Web Appendix 4 of the reference[5]) in the same context.

For fitting random forsest, We used `rfsrc` function from the `randomForestSRC` package, with the following options: `nsplit = 5`, `ntree = 50` and `importance="permute"`. For example: after defining `formula.rF` as the formula object for a given model setting (e.g., `y ~ x`), we can run the random forest model as follows:

```
require(randomForestSRC)
fit <- fsrc(formula = formula.rF, data = admin.data, nsplit = 5,
            ntree = 50, ntime = 10, importance = "permute")
fit$importance
```

R package `Plasmode`[6] provides the `R` functions to simulate plasmode datasets based on user-supplied example studies. We thank the authors of that package (Franklin et al.) for sharing the plasmode simulation implementation codes.

A.7 Plasmode simulation

Healthcare claims databases contain numerous (usually thousands) collected variables. Simulating such a high-dimensional dataset is problematic in a Monte Carlo study because it is difficult to recreate a realistic data generating process that takes into account of associations among a large number of covariates under consideration. Plasmode is a simulation technique that relies on resampling techniques to obtain data that can preserve the empirical associations among the covariates. During the process of plasmode simulation, the analyst can assign a desired value for the true treatment effect in the data generating process. Such a plasmode study begins with an existing cohort, with an assumed data generating process (as in equation (A.3)), and we can modify the existing cohort and injected known effects (signals) into it.

In our study, we used the following outcome generation model for the plasmode simulation:

$$logit\big[Pr(Y=1)\big] = \alpha_0 + \theta \times \alpha_1 T + \gamma \times \alpha_2 X, \tag{A.3}$$

where Y is the outcome (e.g., all-cause mortality following an acute myocardial infarction), T is the treatment indicator (whether or not the patient being treated with statin), X is the high-dimensional covariate matrix that includes the important investigator-specified covariates (listed in eTable A.1), additional investigator-specified covariates (listed in eTable A.2) and the list of created empirical covariates obtained by running the hdPS algorithm on the complete statin user dataset with $32,792$ patients. These empirical variables should act as proxy or surrogate of the unmeasured confounders. As for the parameters in equation (A.3), α_0 is the intercept, α_1 is the treatment effect, α_2 is the vector of effects associated with covariates listed in X, θ is the treatment effect multiplier and γ is the covariate effect multiplier.

From the above outcome generation model, in each of 18 simulation scenarios considered in this study, we have generated $N = 500$ datasets each with $m = 10,000$ patients. Note that, for each of these newly generated datasets (with $10,000$ patients), we have separately prioritized the empirical covariates by applying the hdPS algorithm on each of these datasets[5,7]. Therefore, the top 500 hdPS variables for a given dataset may not be identical to those obtained from another dataset. The variation in the resulting effect measures (RD or OR) from different datasets comes not only from the differences in hdPS variables in each dataset but also from the resampling procedure (i.e., selection of $10,000$ patients with replacement out of $32,792$ patients) integrated in the plasmode simulation algorithm.

The plasmode simulation algorithm samples exposed and unexposed subjects with replacement from the empirical dataset in such a way that guarantees a desired study size (m) and a prevalence of exposure (p_E) in the simulated plasmode samples[5,7,8]. Also, this simulation algorithm allows researchers to specify the intercept value in the outcome-generating model to guarantee a desired prevalence of outcome (p_Y)[5,7].

Methodologically, the plasmode simulation realistically generates the data by controlling the relationship with outcome by retaining α_2 estimates (parameter estimates associates with the covariates) in the outcome generation model (equation (A.3)) same as the estimates obtained from the empirical data fitting. The plasmode simulation uses resampling techniques such as bootstrap to select patients in a specific sample with replacement. Here, the bootstrap samples (of specified size m) are collected from the complete set of covariate-exposure matrix $Z = (T, X)$. As none of these variables in the covariate-exposure matrix, Z are permuted or modified in any way, in each bootstrap sample (of a reasonable size), systematically, the relationships should remain intact among exposure and covariates[7]. Therefore, relationship with covariates and outcomes are controlled by fixing α_2 values in the outcome generation model and boostrap ensures joint distribution of exposure and covariates are unaltered, there should not be any obvious reason why the relationship among covariates and exposure should be different in plasmode samples. In that sense, in the plasmode simulation, the 'amount of confounding' from a covariate (i.e., relationship of a covariate with the outcome as well as the exposure; both of which relationships are required for a covariate to be considered as a confounder) is controlled[7].

However, among other things, this simulation mechanism do allow researchers to change the multipliers of the treatment effect and the covariate effects by changing θ parameter value and

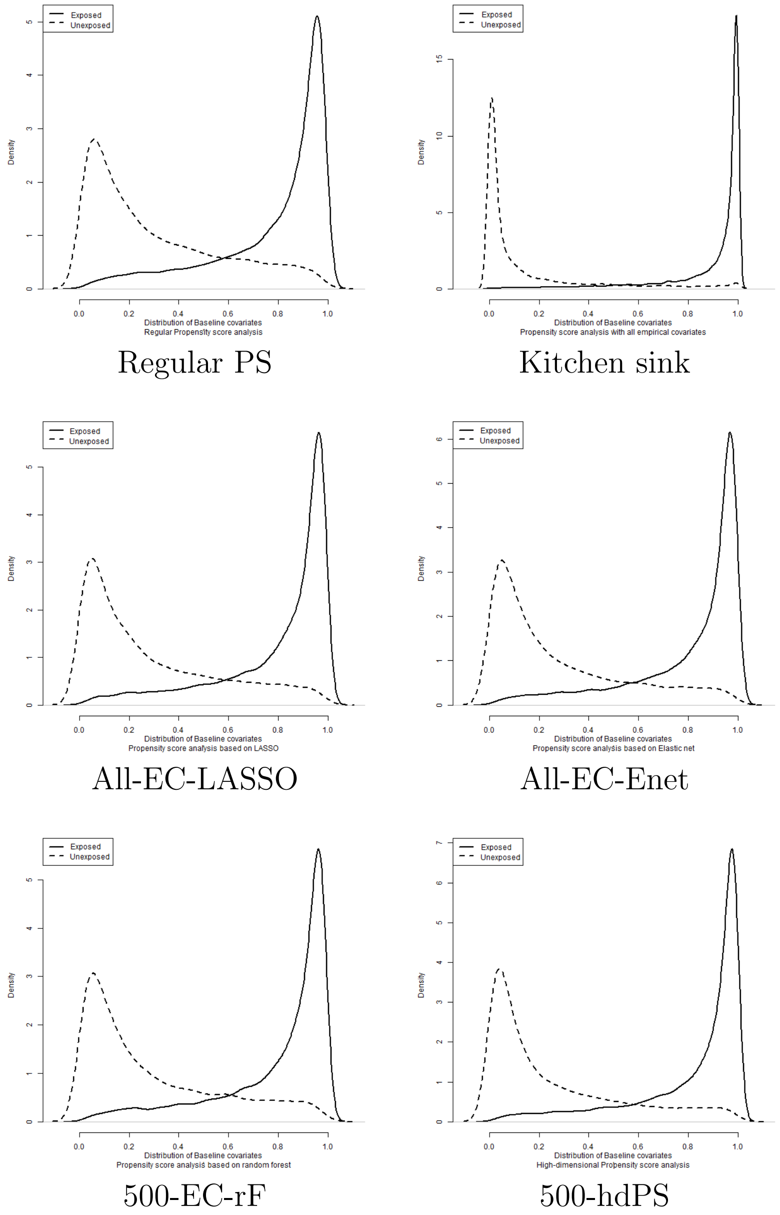
γ parameter vector respectively. In certain combination of these parameters values, it is possible that an important confounder in the empirical study may not remain important in the plasmode samples. Future research should investigate further in this issue.

Note that, the important confounders (age, sex, obesity, smoking, and history of diabetes) considered in this study were not based on their higher strength of association with outcome and exposure in the empirical data, but based on subject-specific knowledge from previous research[9]. The idea of the sensitivity analysis done in our study was not to see the impact of excluding covariates that were highly association with the outcome and the exposure (e.g., strong confounders), but to see if hdPS algorithm can account for useful information that are not collected during data collection stage by using proxy data (empirical covariates). Instead of making up new covariates, we have decided to delete some real covariates that were considered useful by the experts[9].

Plasmode simulations are built based on a given empirical data setting, and the generalizability of the results is an issue for such simulations. To convince the users and the analysts, more such plasmode simulations mimicking other healthcare administrative datasets should be conducted to validate various machine-learning and hybrid methods.

A.8 Balance diagnostics and Data Analyses

A.8.1 Balance diagnostics



Regular PS

Kitchen sink

All-EC-LASSO

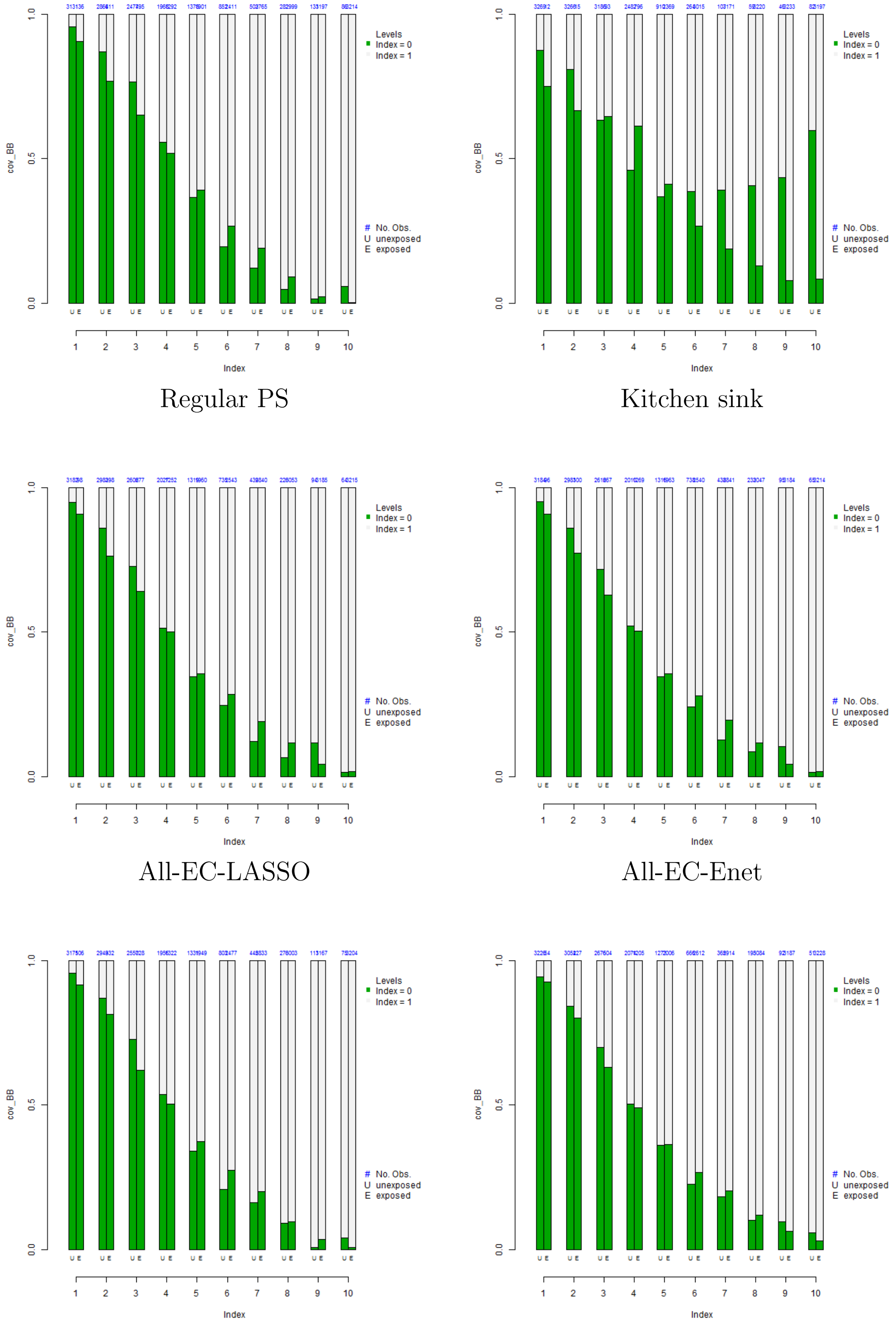
All-EC-Enet

500-EC-rF

500-hdPS

eFigure A.4: Balance

For the purpose of illustration, we checked the balance of the beta-blocker covariate, and we observe that there are imbalances in the last few deciles of PSs when we considered all empirical-covariates. However, when we selected the 500 top ranked hdPS variables, the balance is regained (see eFigure A.5).



Regular PS

Kitchen sink

All-EC-LASSO

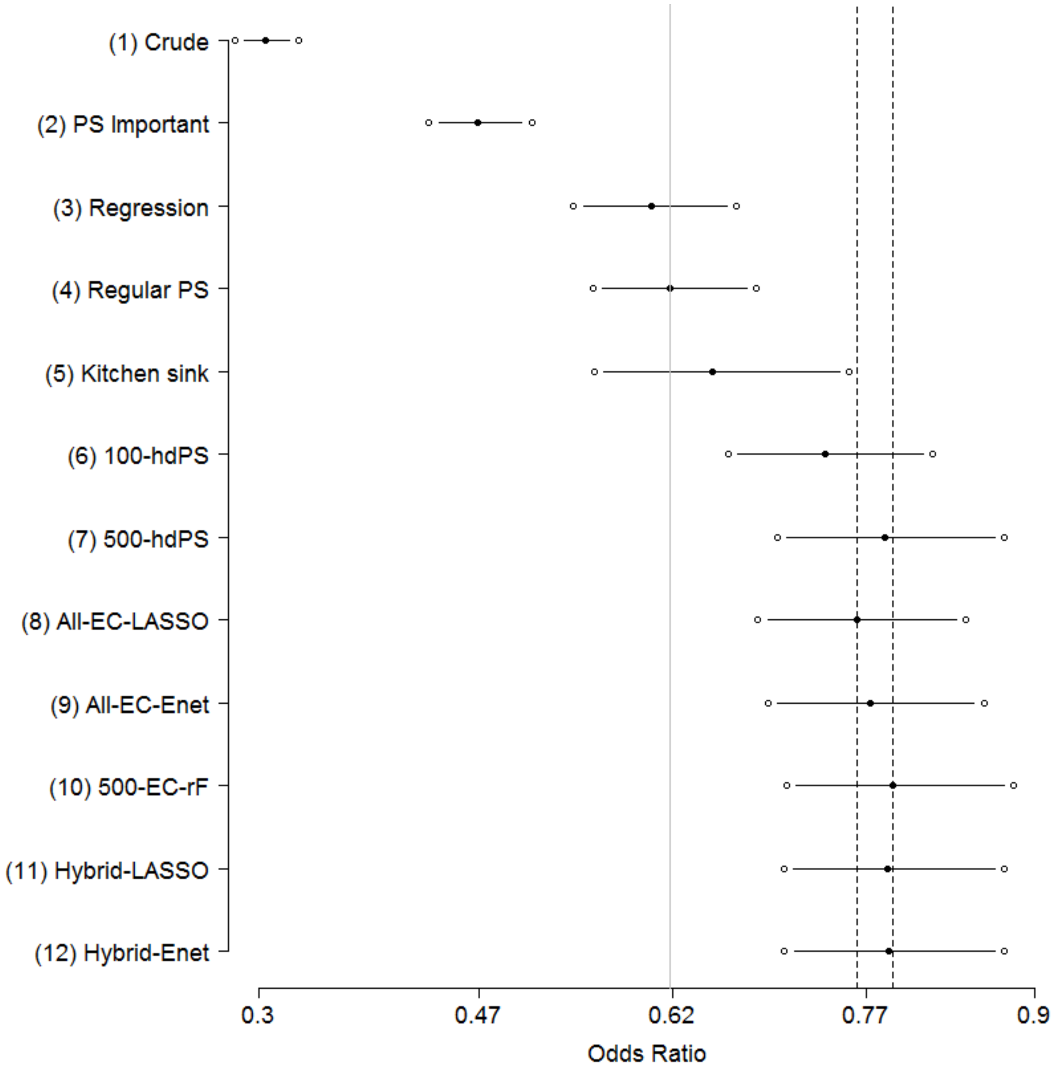
All-EC-Enet

500-EC-rF

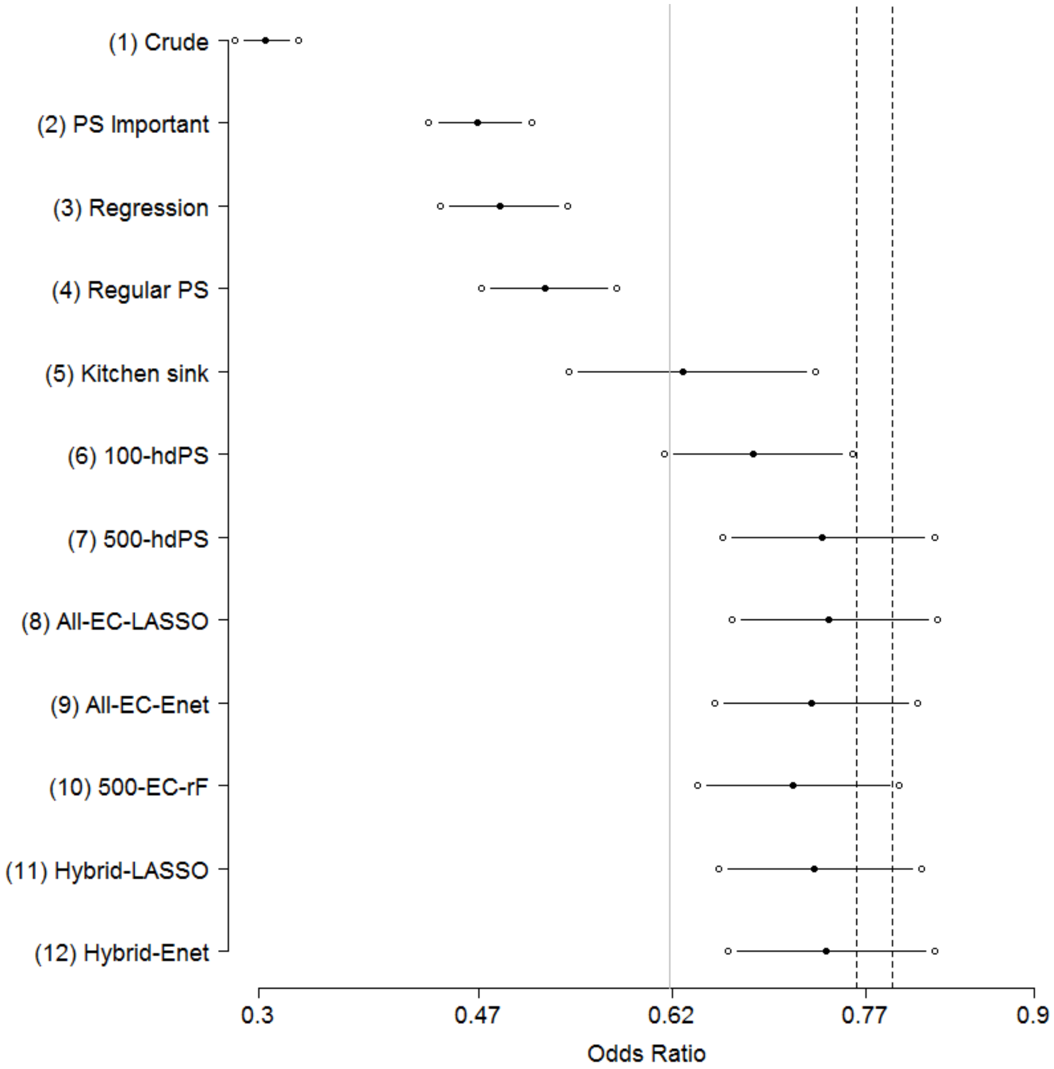
500-hdPS

eFigure A.5: Balance for beta clocker

A.8.2 Data Analyses



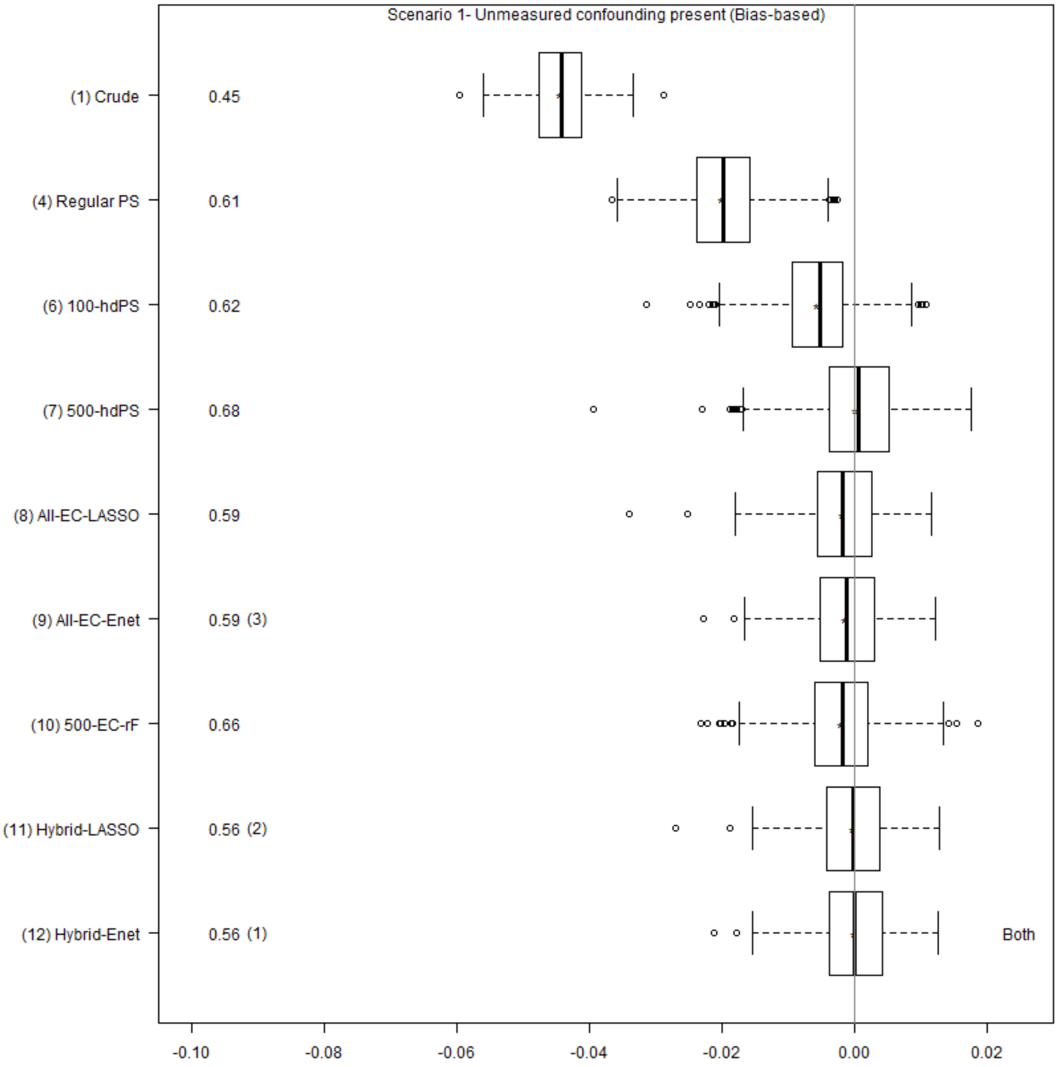
eFigure A.6: Analysis results from the approaches under consideration. When only the investigator-specified covariates were considered, the estimated OR was 0.62 in our analysis (represented by the solid grey line). When considering 500 or more empirical covariates and all the investigator-specified covariates in the analysis, the estimated ORs were between 0.76 and 0.79 in our analysis (represented by the dotted lines). Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; OR, odds ratio; EC, empirical covariate.



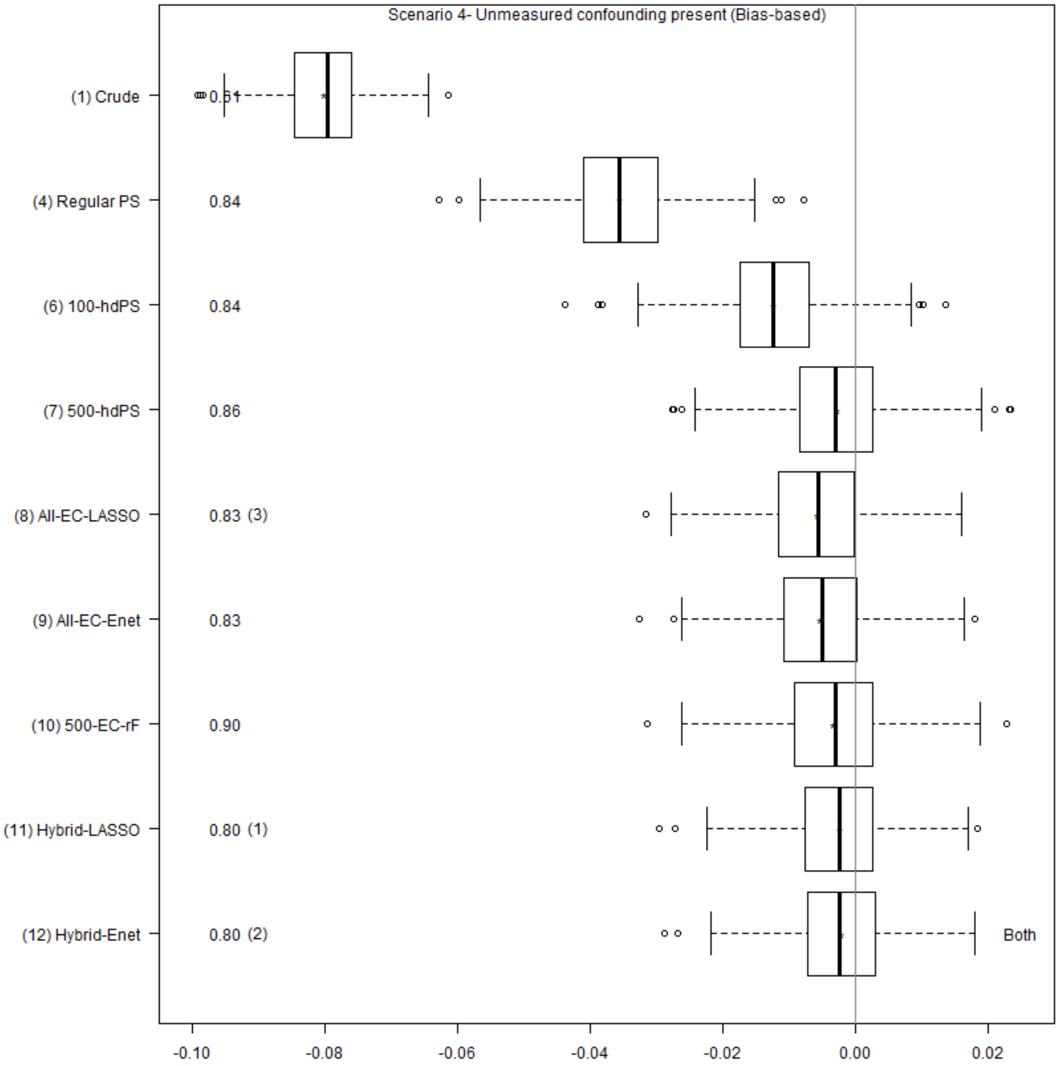
eFigure A.7: High-dimensional propensity score and machine learning alternative results without the five important covariates: age, sex, obesity, smoking, and history of diabetes. For comparison with the analyses with these five covariates, the solid grey line represents the estimated OR of 0.62 (when all the investigator-specified covariates were considered in our analysis), and the dotted lines represent the estimated ORs 0.76 and 0.79 (the range of estimated ORs, when considering 500 or more empirical covariates in the analysis including all the investigator-specified covariates in our analysis.) Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; OR, odds ratio; EC, empirical covariate.

A.9 Figures from Plasmode simulation

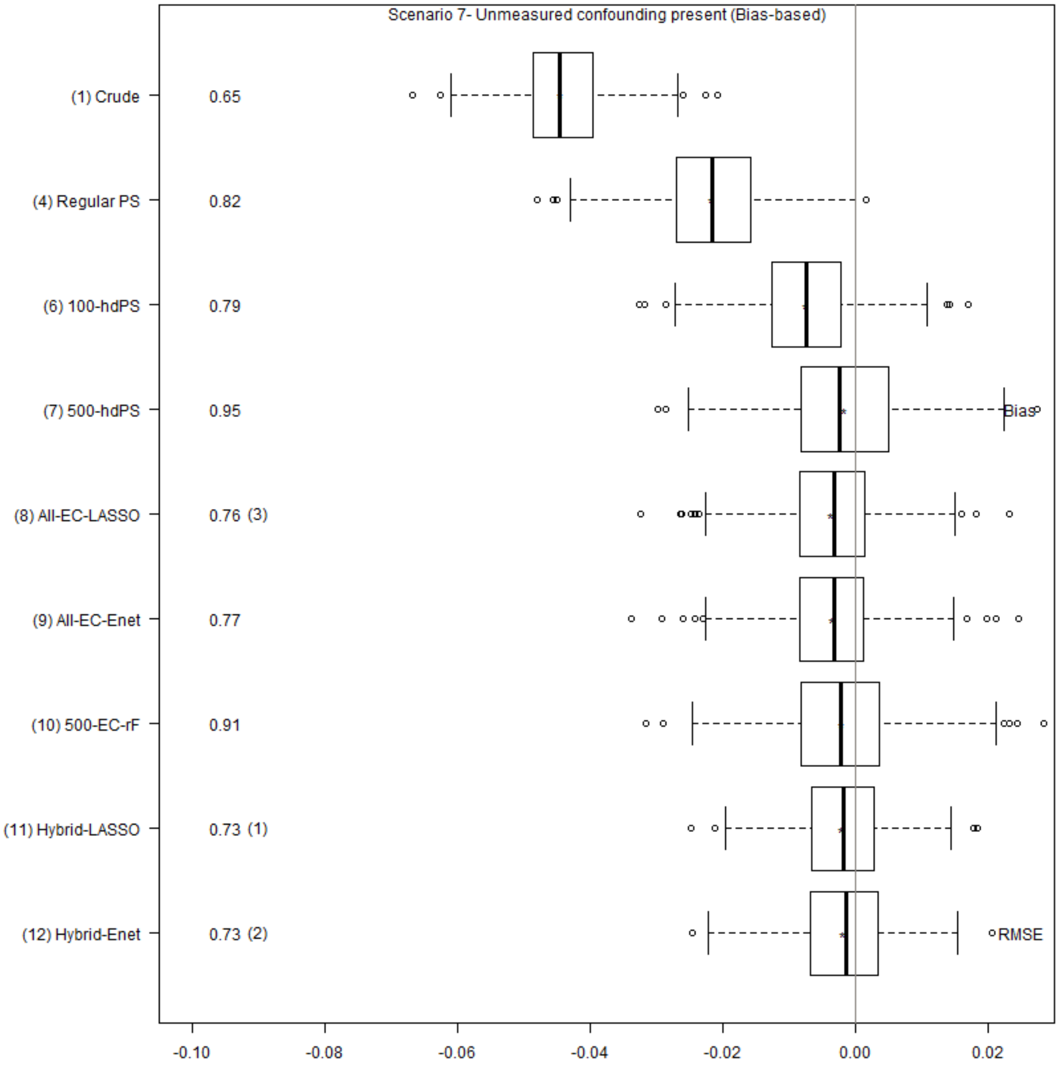
A.9.1 Unmeasured confounding present (Bias-based analysis): Main three scenarios:



eFigure A.8: Side-by-side boxplots of the estimated risk differences (from 500 datasets) via the approaches under consideration in the plasmode Simulation Scenario 1-U. Corresponding mean values are marked by $*$. The indicator "Both" means the approach is found best by both MSE and bias criteria. Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; RD, risk difference; EC, empirical covariate.

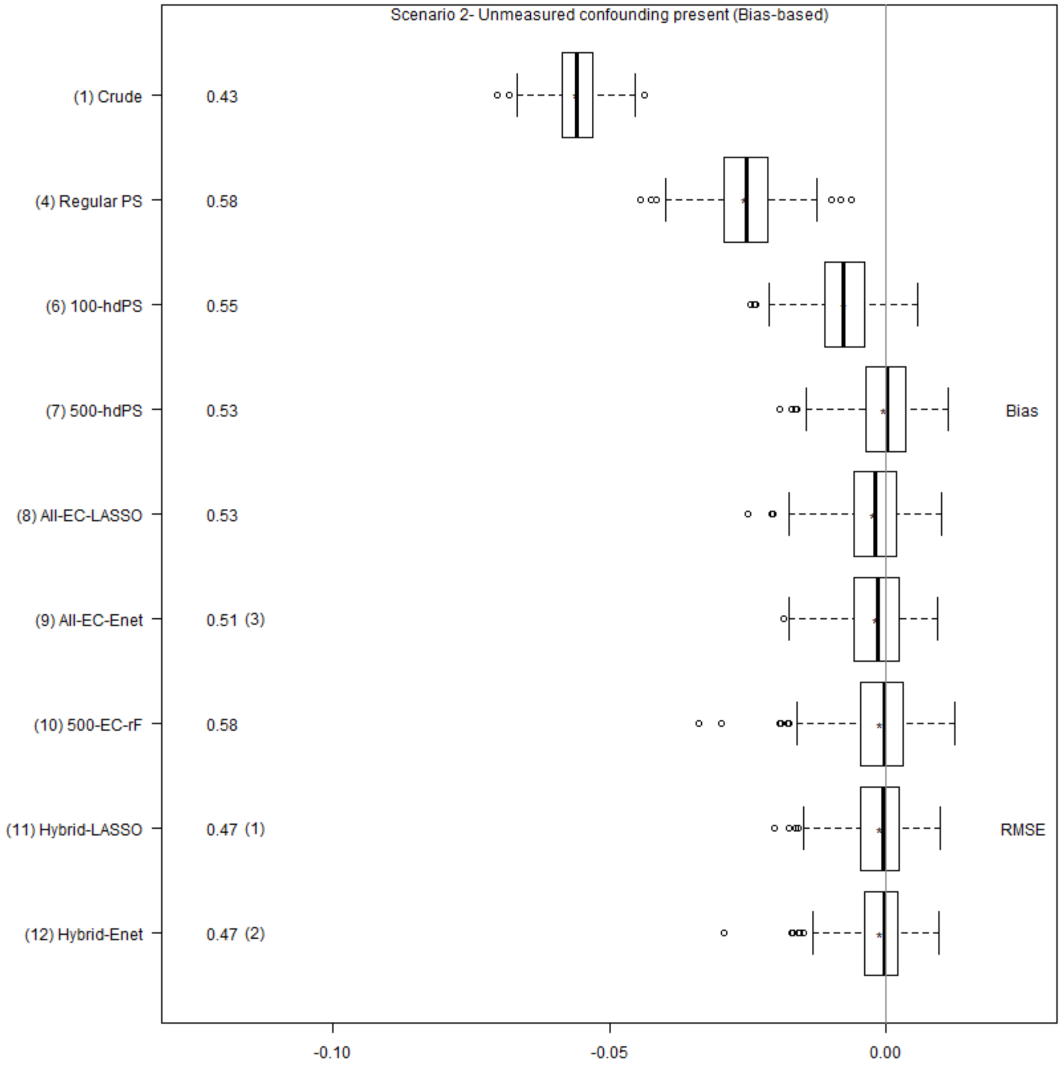


eFigure A.9: Side-by-side boxplots of the estimated risk differences (from 500 datasets) via the approaches under consideration in the plasmode Simulation Scenario 4-U. Corresponding mean values are marked by ∗. The indicator “Both” means the approach is found best by both MSE and bias criteria. Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; RD, risk difference; EC, empirical covariate.

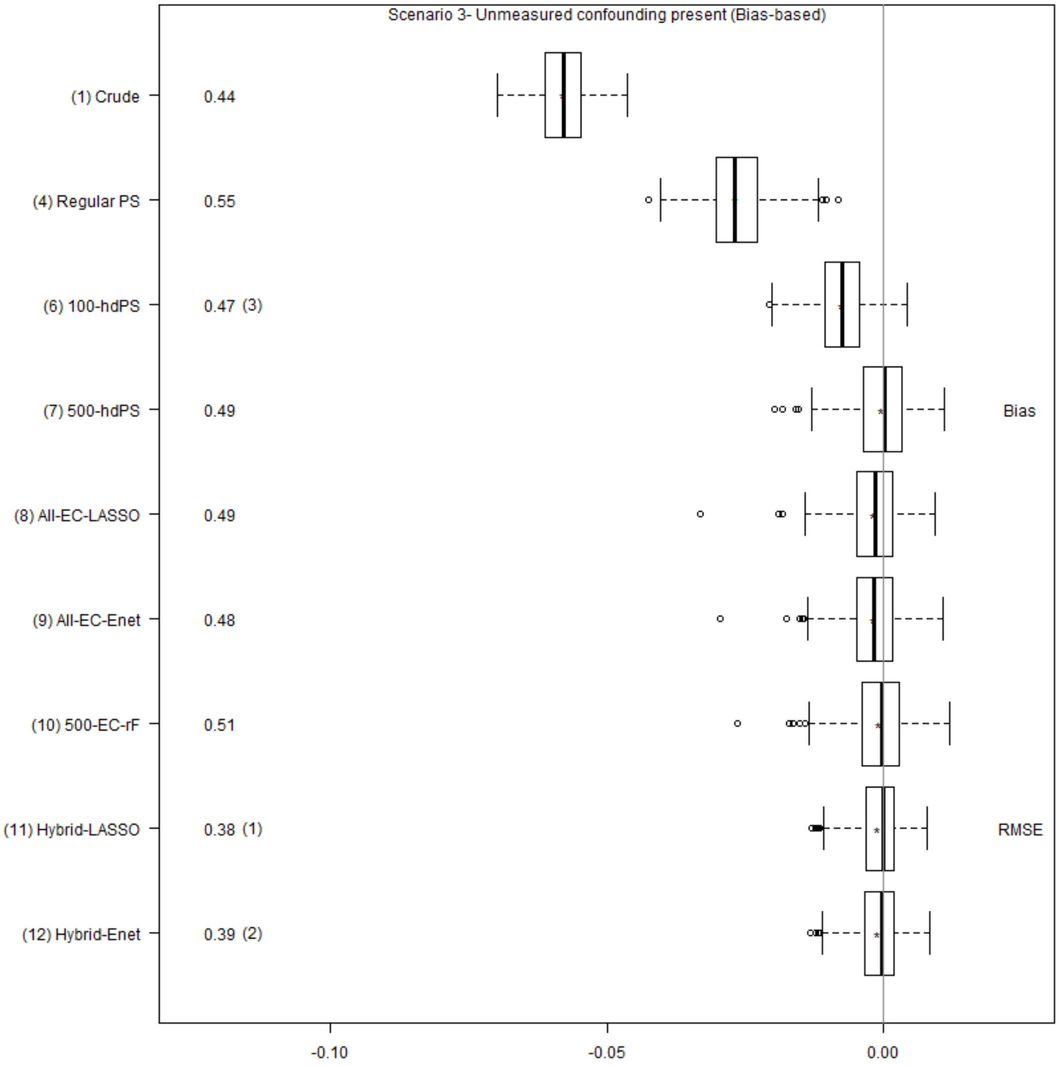


eFigure A.10: Side-by-side boxplots of the estimated risk differences (from 500 datasets) via the approaches under consideration in the plasmode Simulation Scenario 7-U. Corresponding mean values are marked by $*$. The indicator "RMSE" means the approach is found best by the RMSE criterion and the indicator "Bias" means the approach is found best by the bias criterion. Abbreviations: PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; RD, risk difference; EC, empirical covariate; RMSE, root mean squared error.

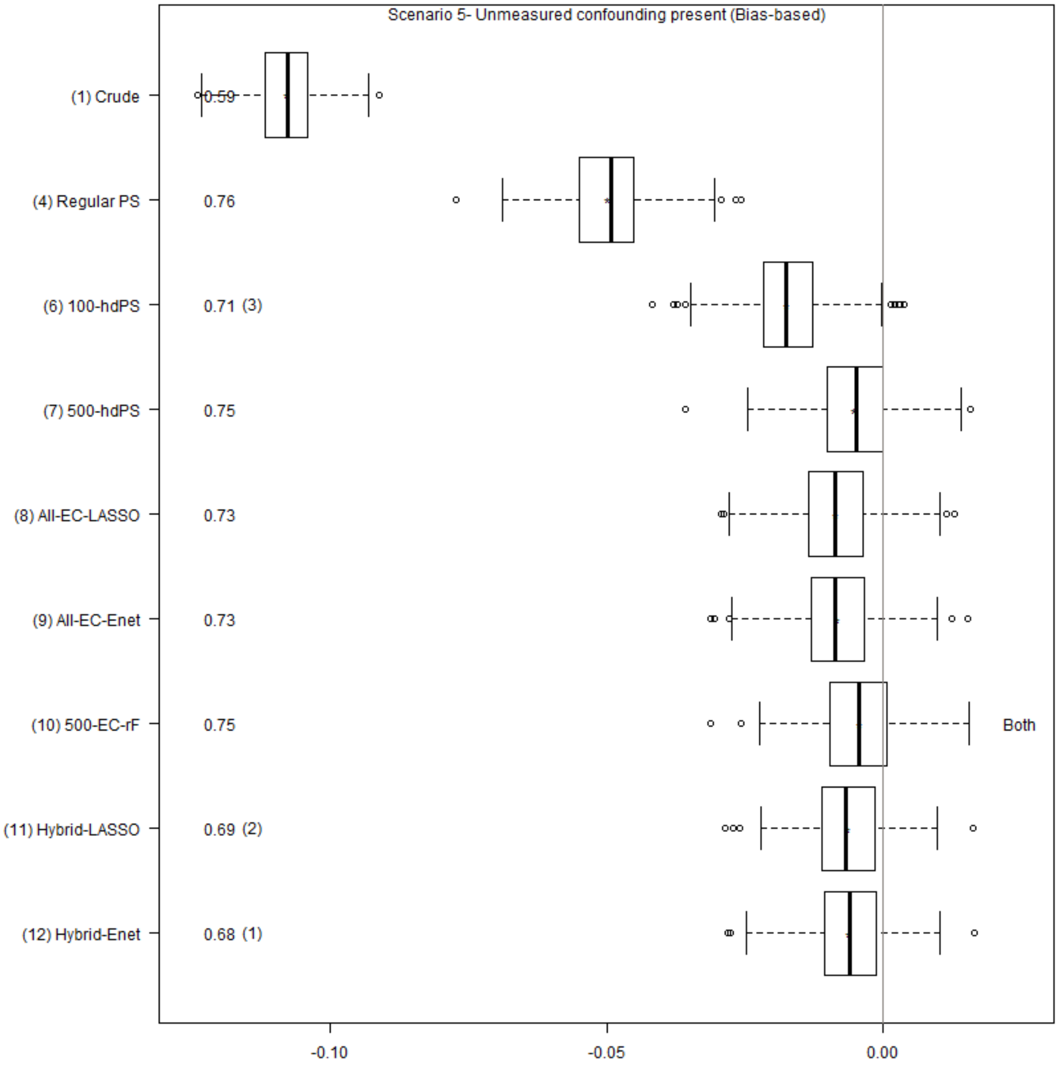
A.9.2 Unmeasured confounding present (Bias-based analysis): Other scenarios:



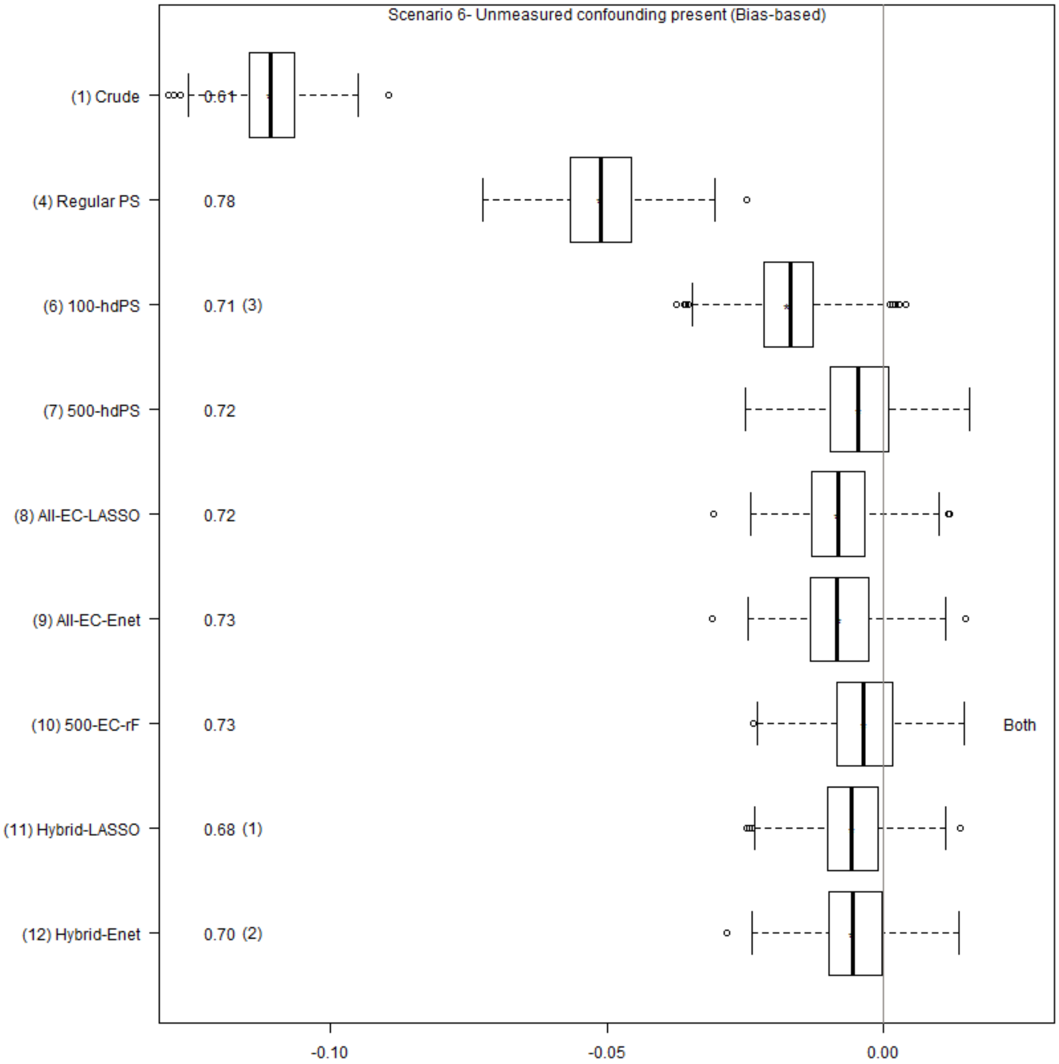
eFigure A.11: Plasmode Simulation Scenario 2-U



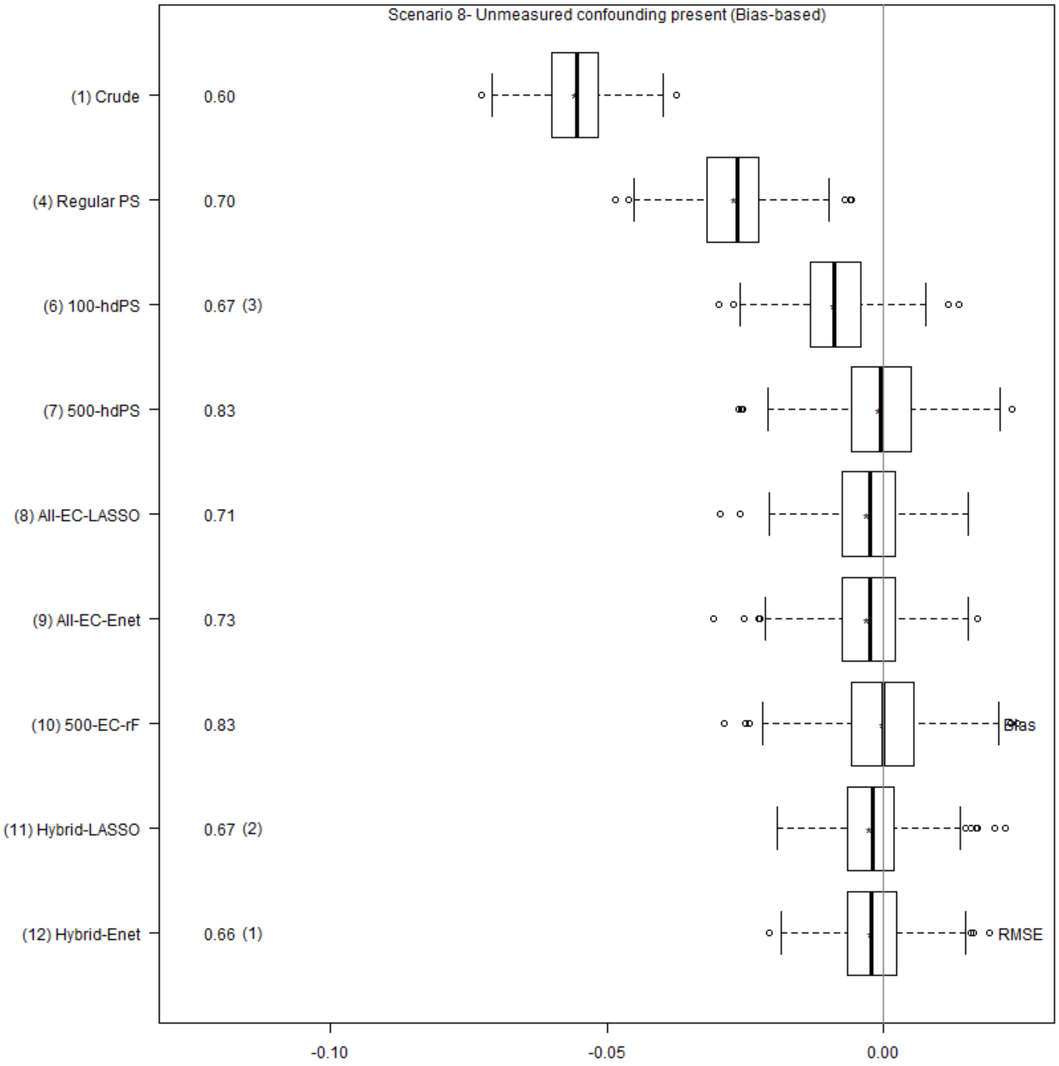
eFigure A.12: Plasmode Simulation Scenario 3-U



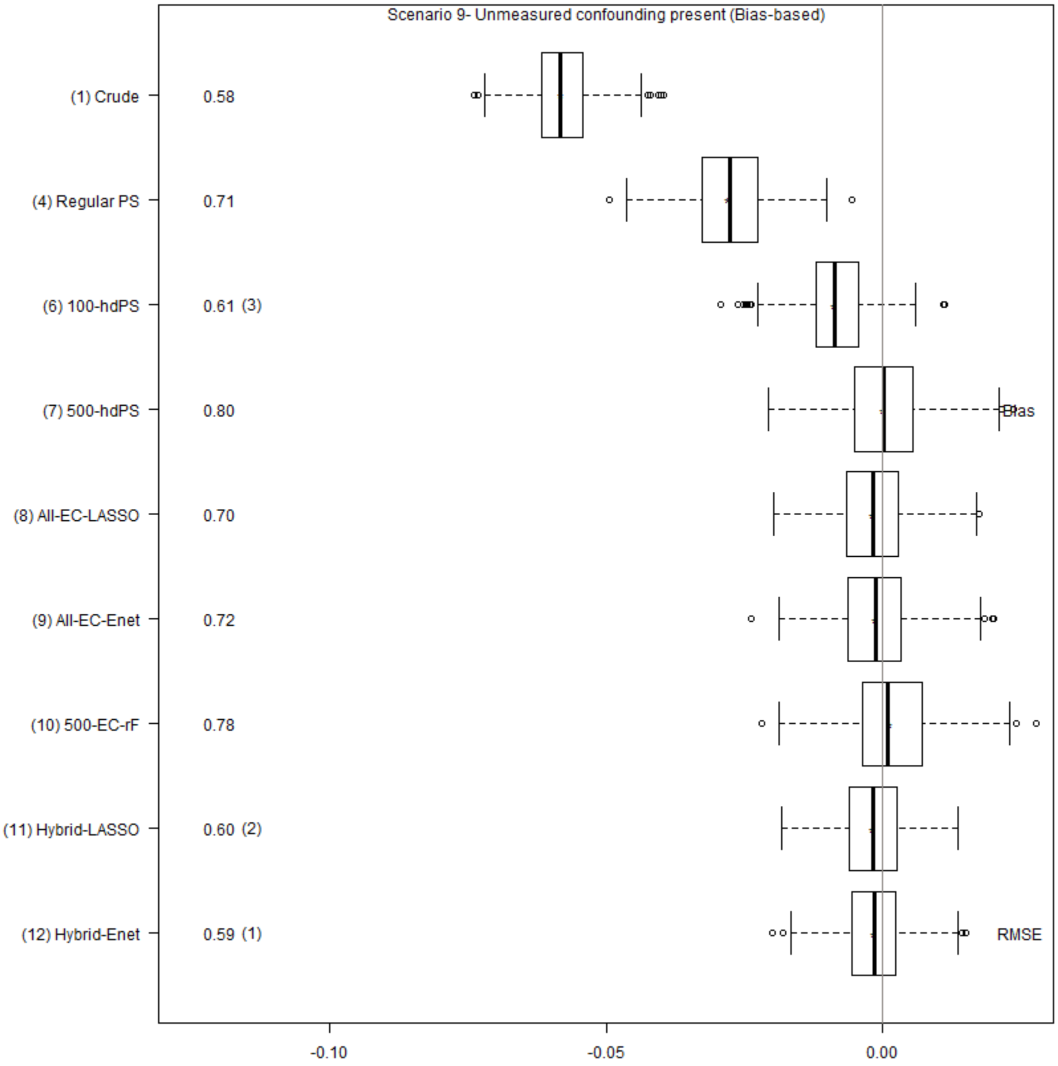
eFigure A.13: Plasmode Simulation Scenario 5-U



eFigure A.14: Plasmode Simulation Scenario 6-U

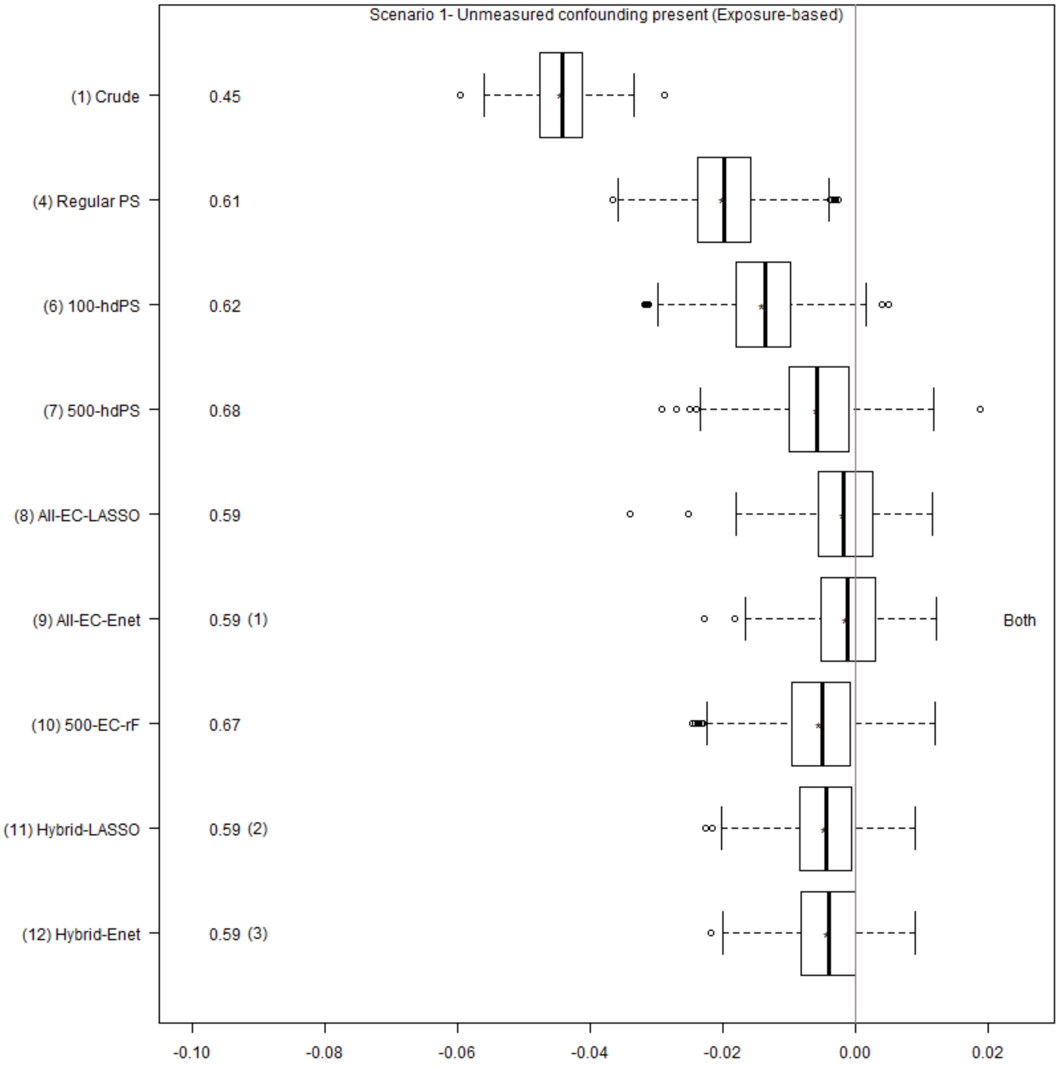


eFigure A.15: Plasmode Simulation Scenario 8-U

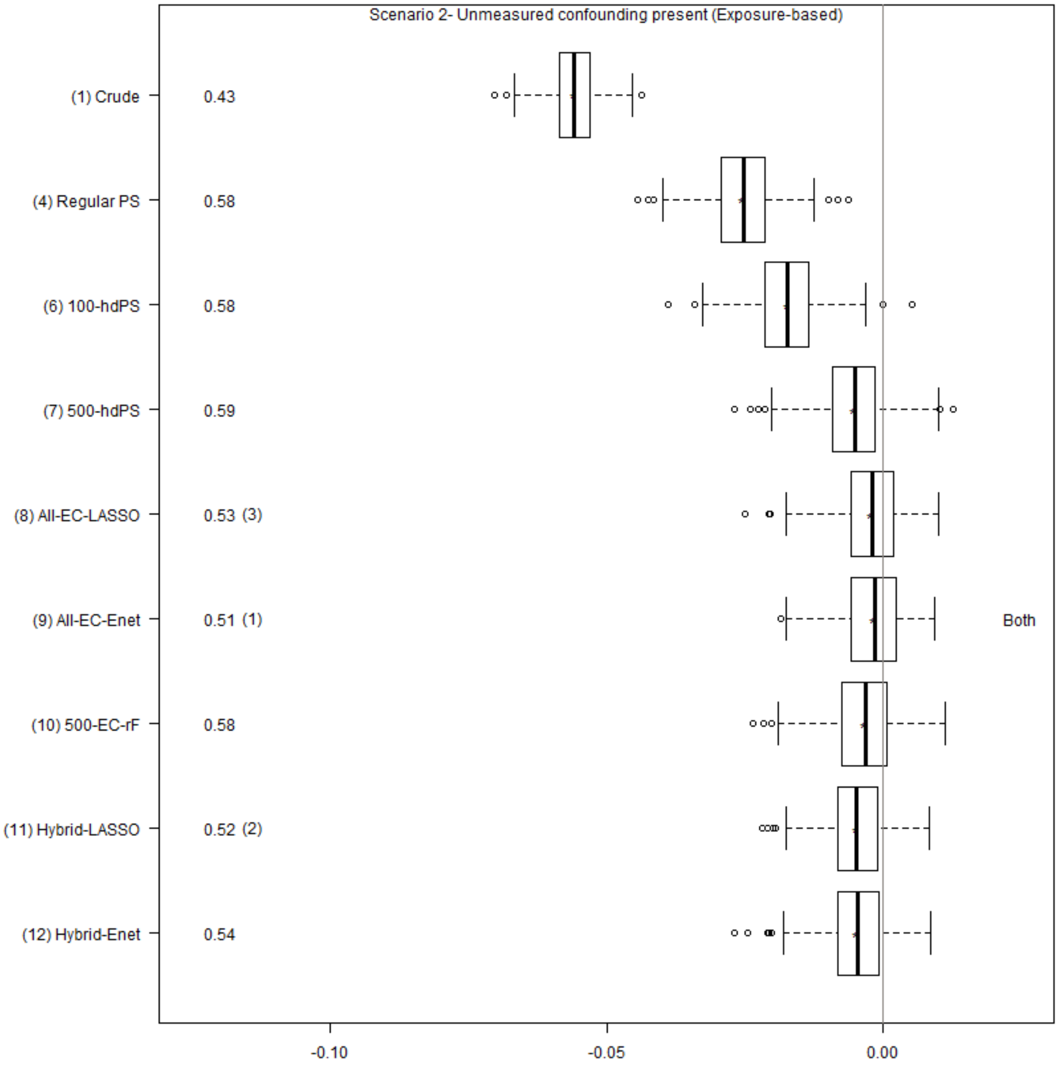


eFigure A.16: Plasmode Simulation Scenario 9-U

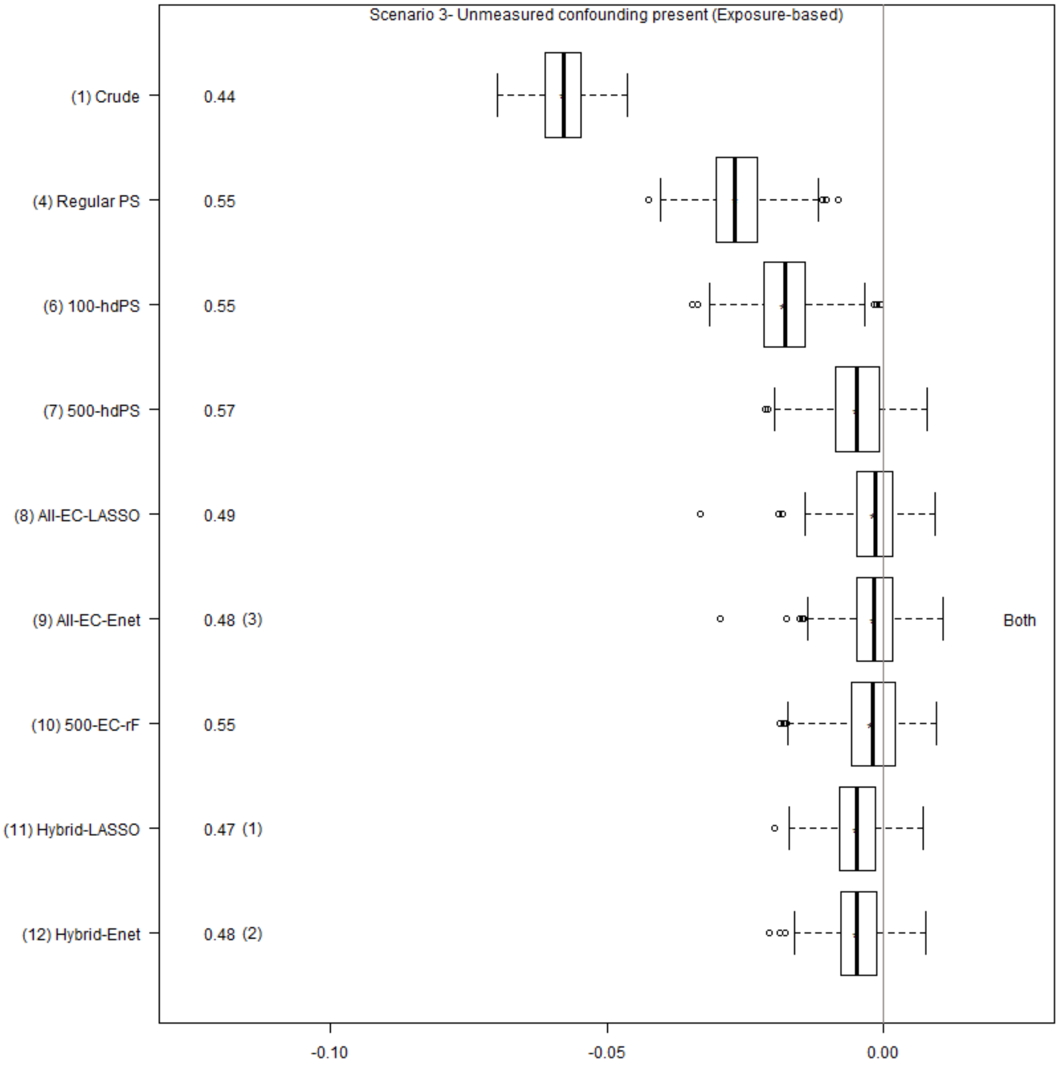
A.9.3 If unmeasured confounding present (Exposure-based analysis)



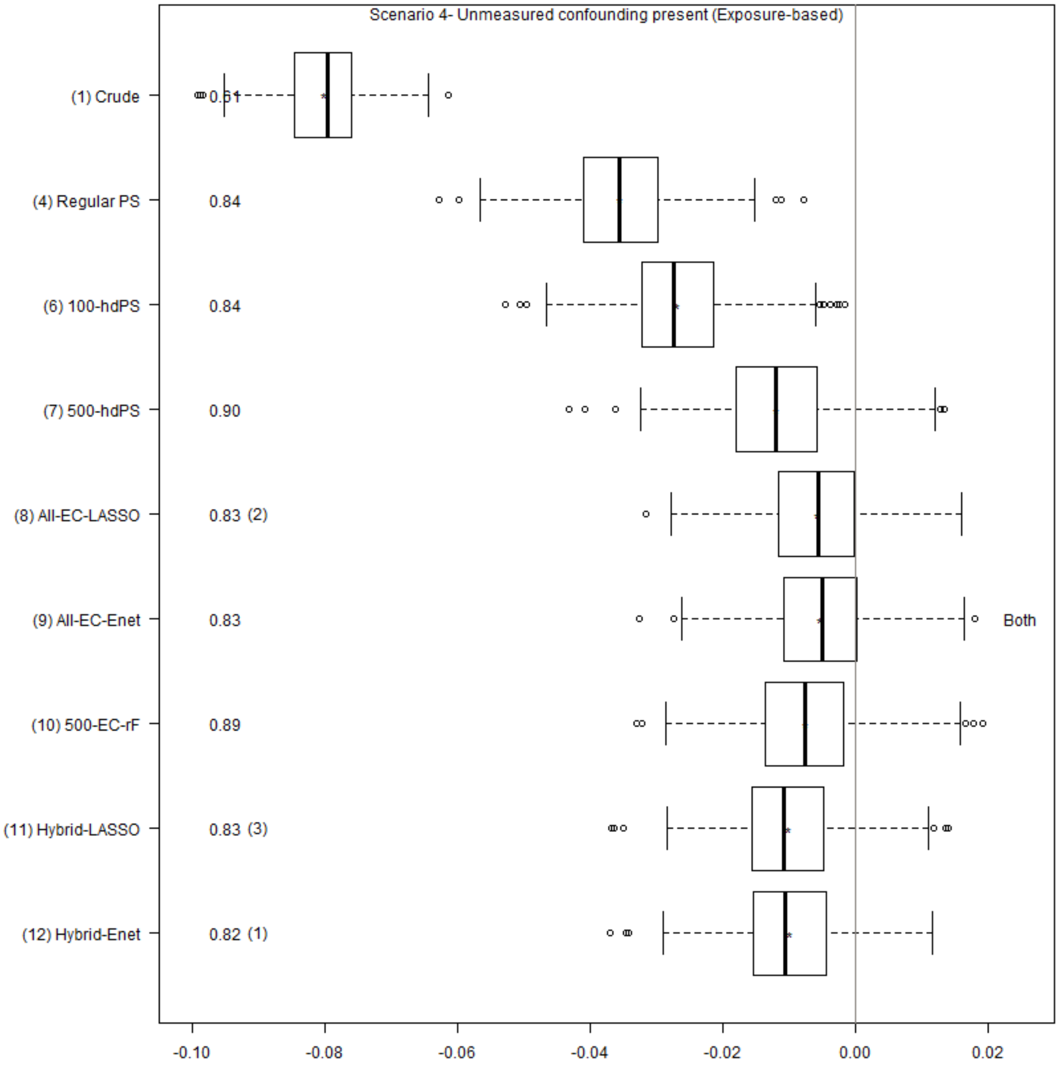
eFigure A.17: Plasmode Simulation Scenario 1-A



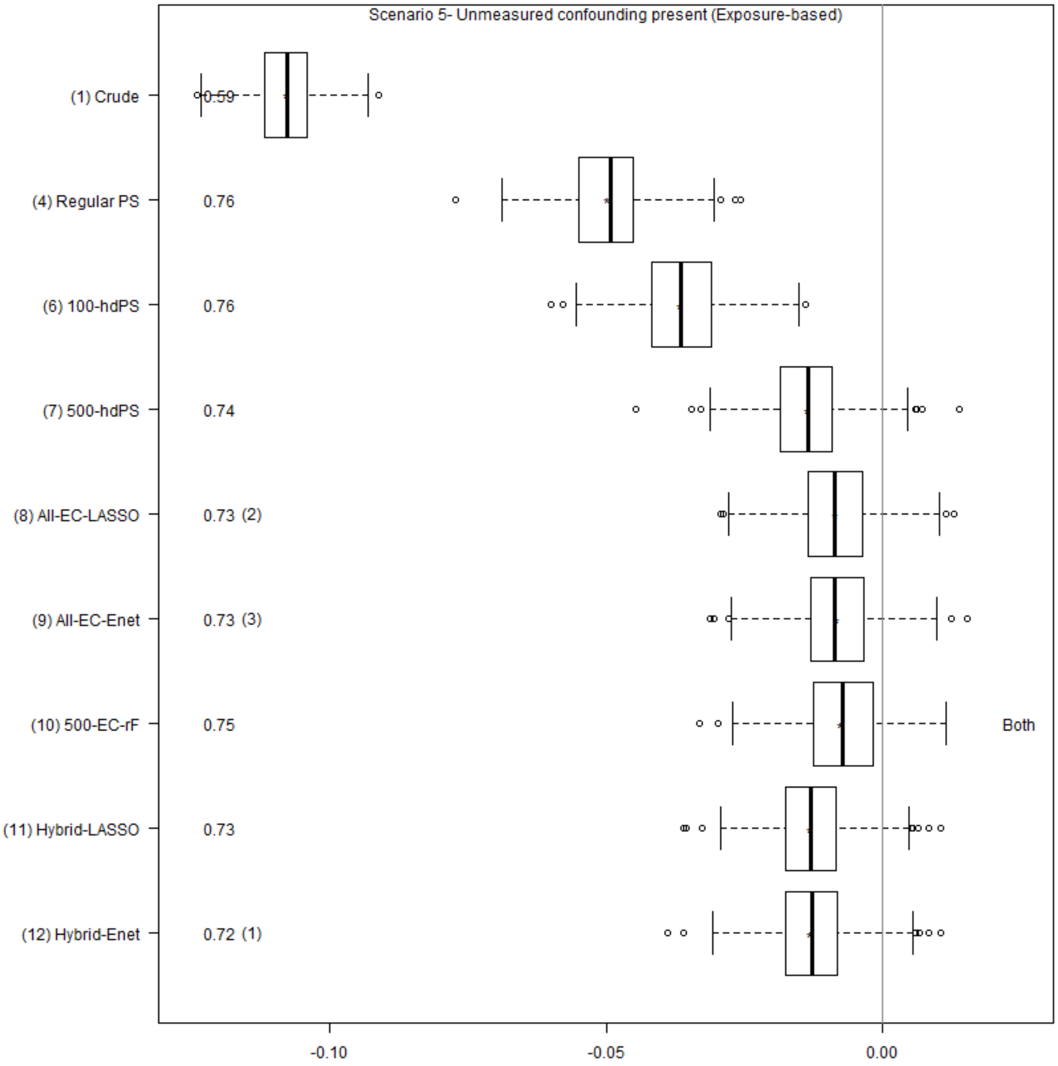
eFigure A.18: Plasmode Simulation Scenario 2-A



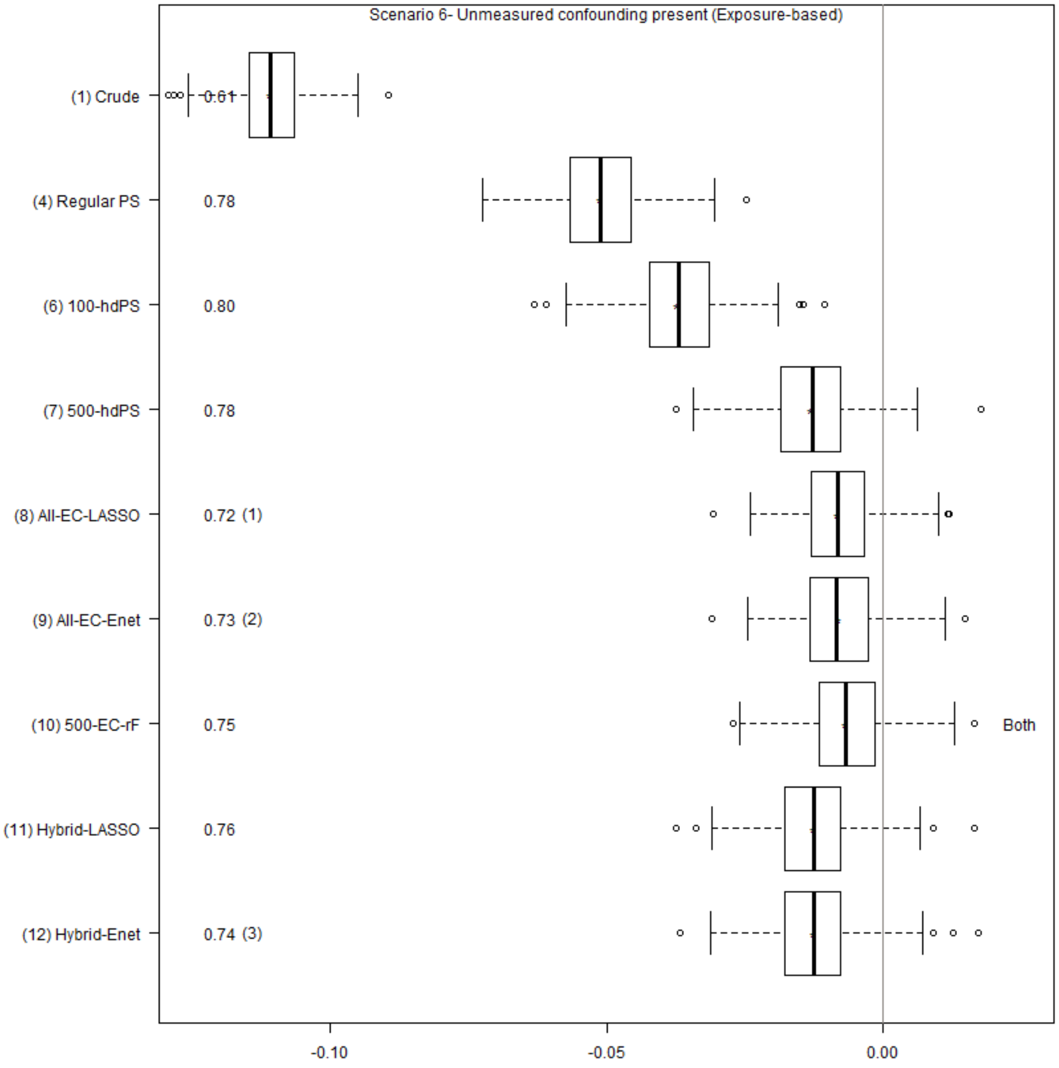
eFigure A.19: Plasmode Simulation Scenario 3-A



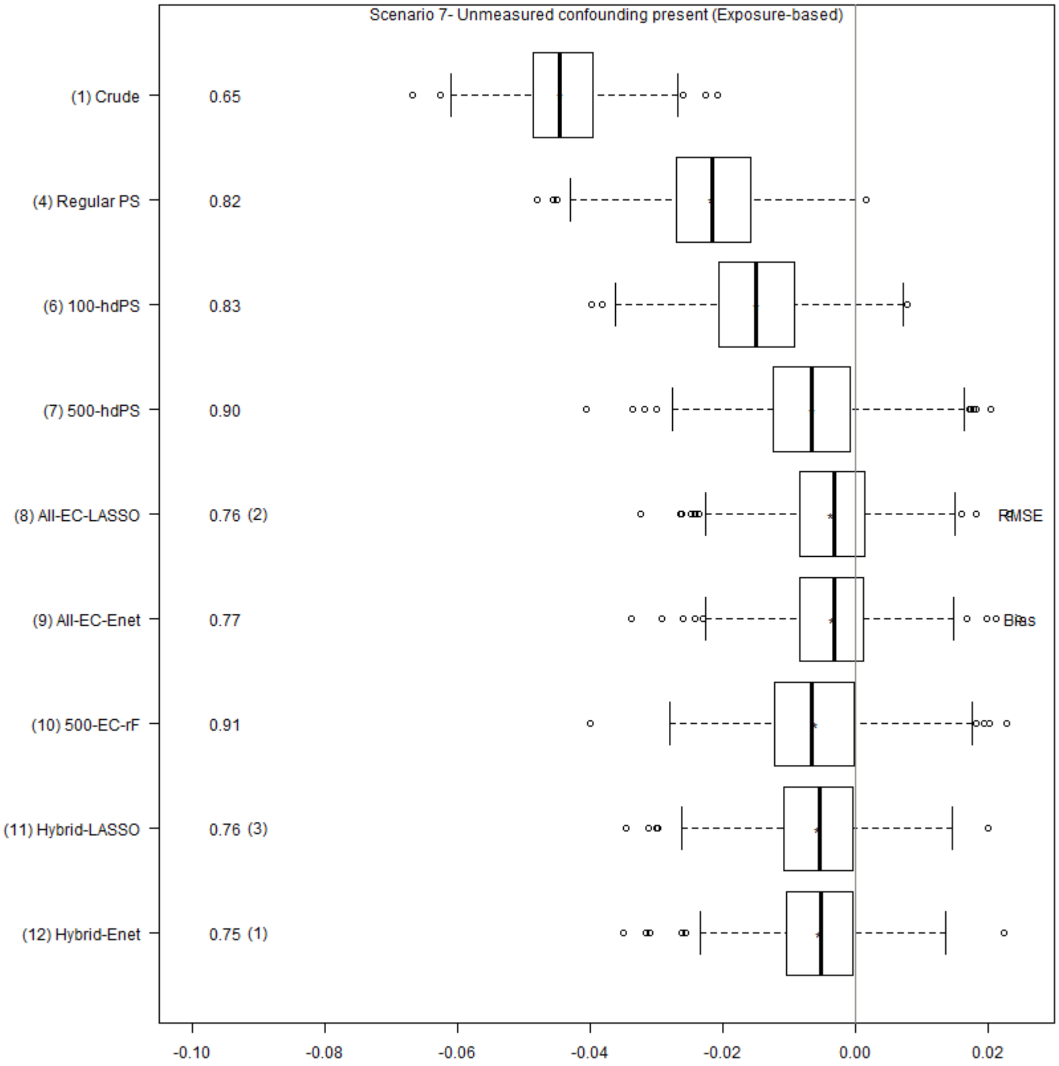
eFigure A.20: Plasmode Simulation Scenario 4-A



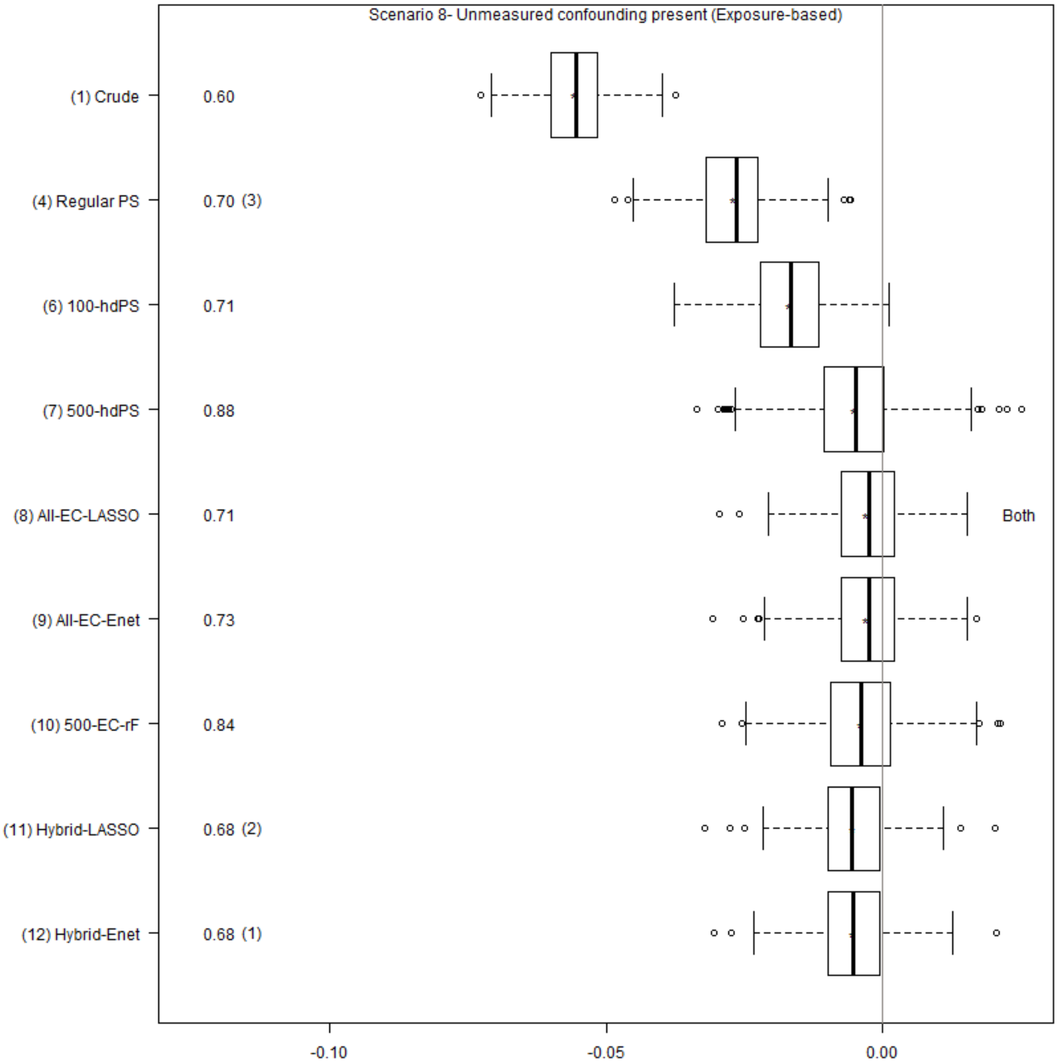
eFigure A.21: Plasmode Simulation Scenario 5-A



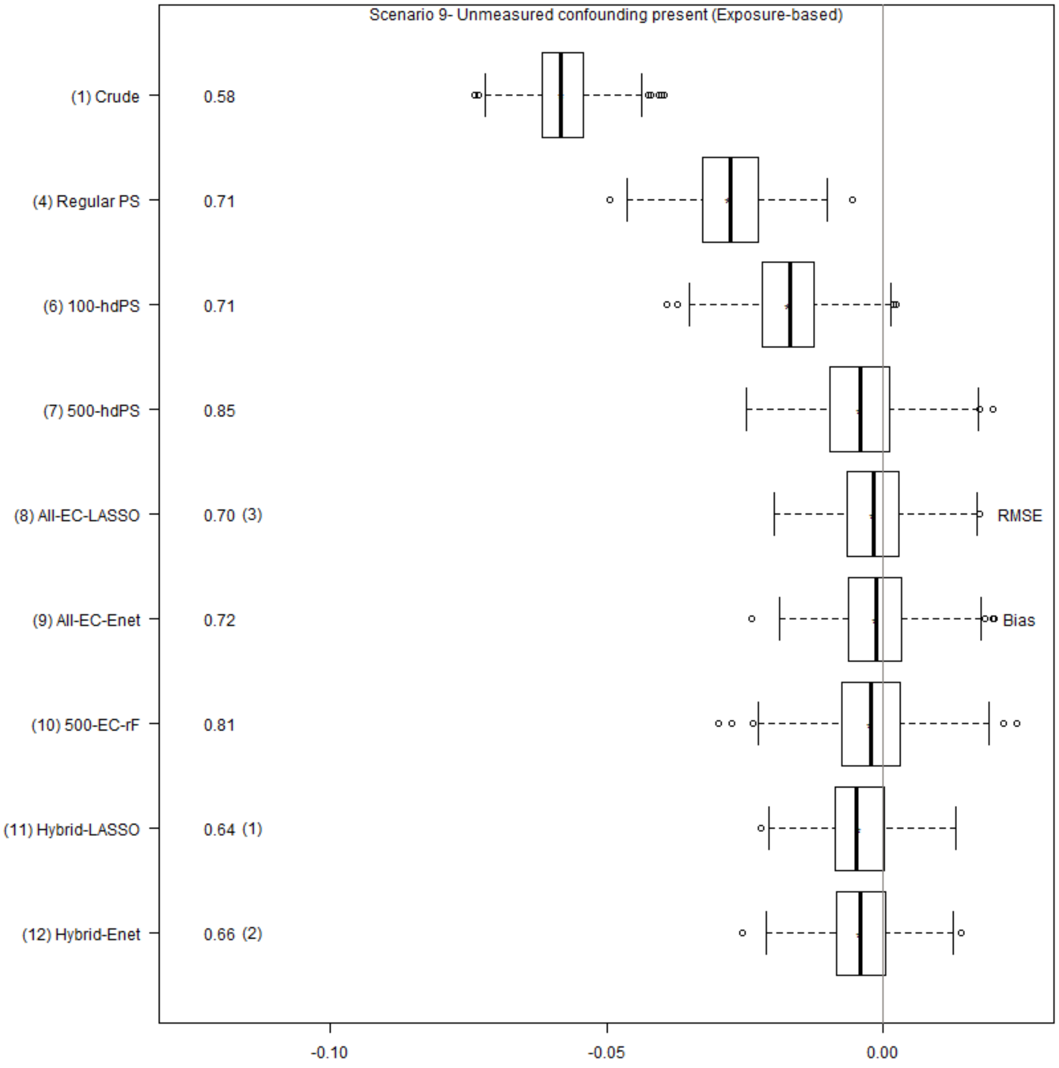
eFigure A.22: Plasmode Simulation Scenario 6-A



eFigure A.23: Plasmode Simulation Scenario 7-A

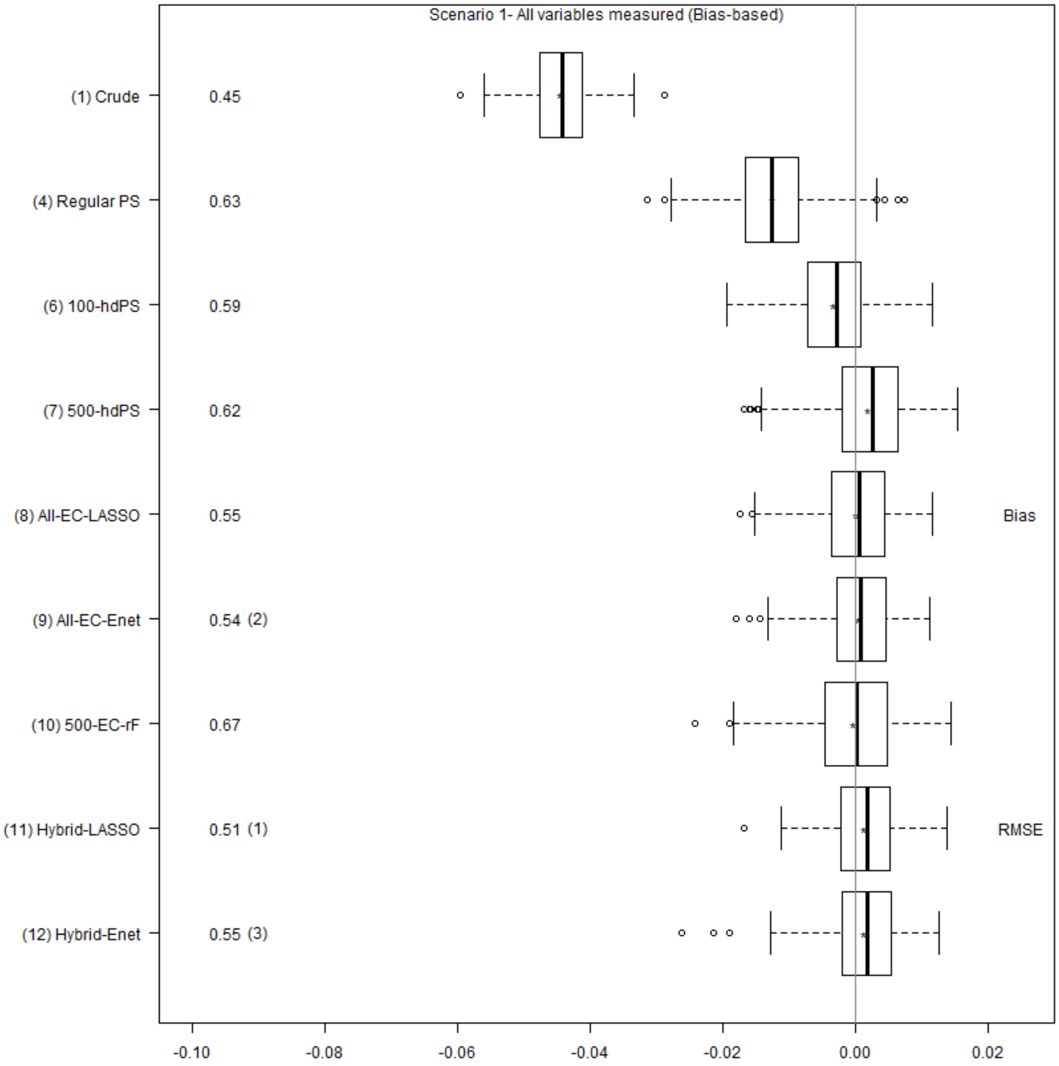


eFigure A.24: Plasmode Simulation Scenario 8-A

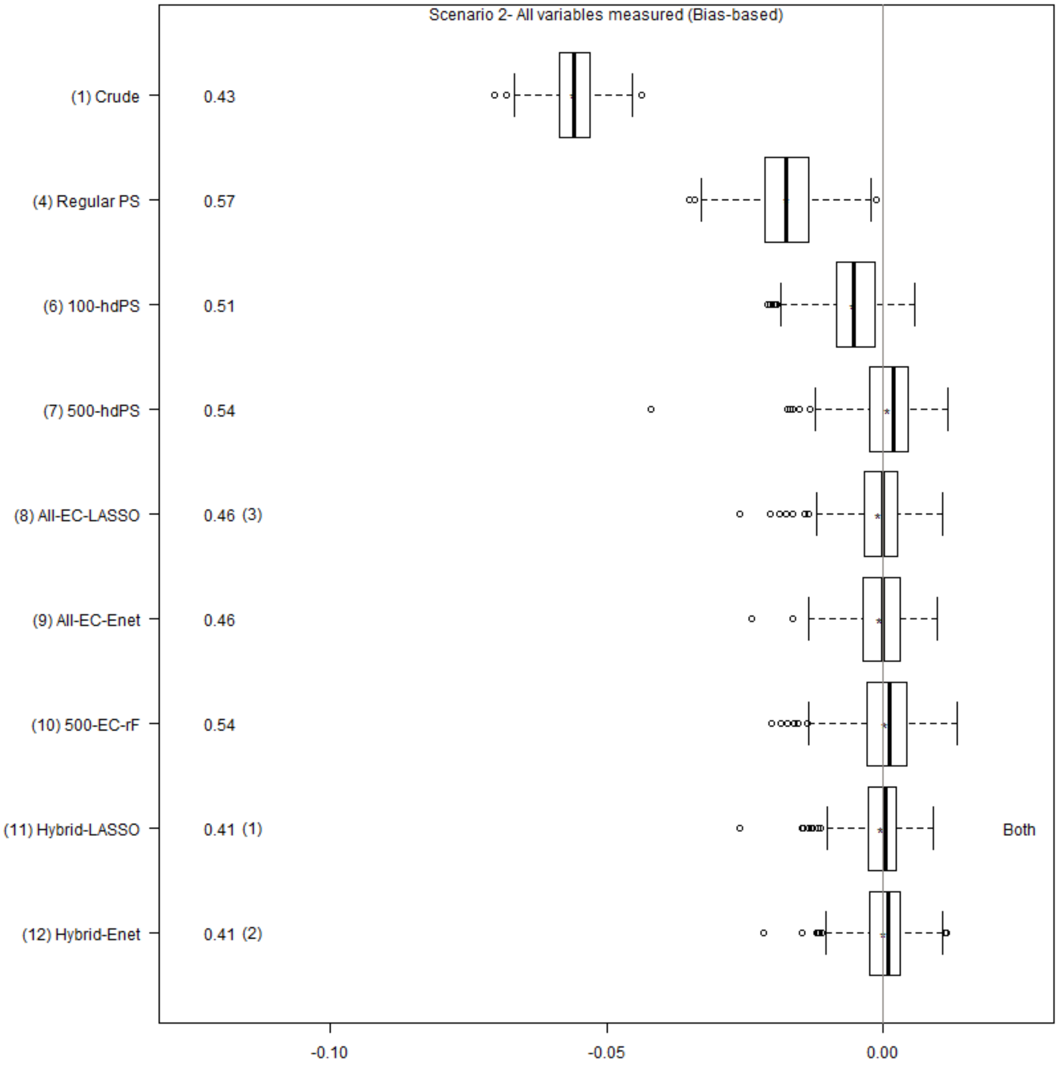


eFigure A.25: Plasmode Simulation Scenario 9-A

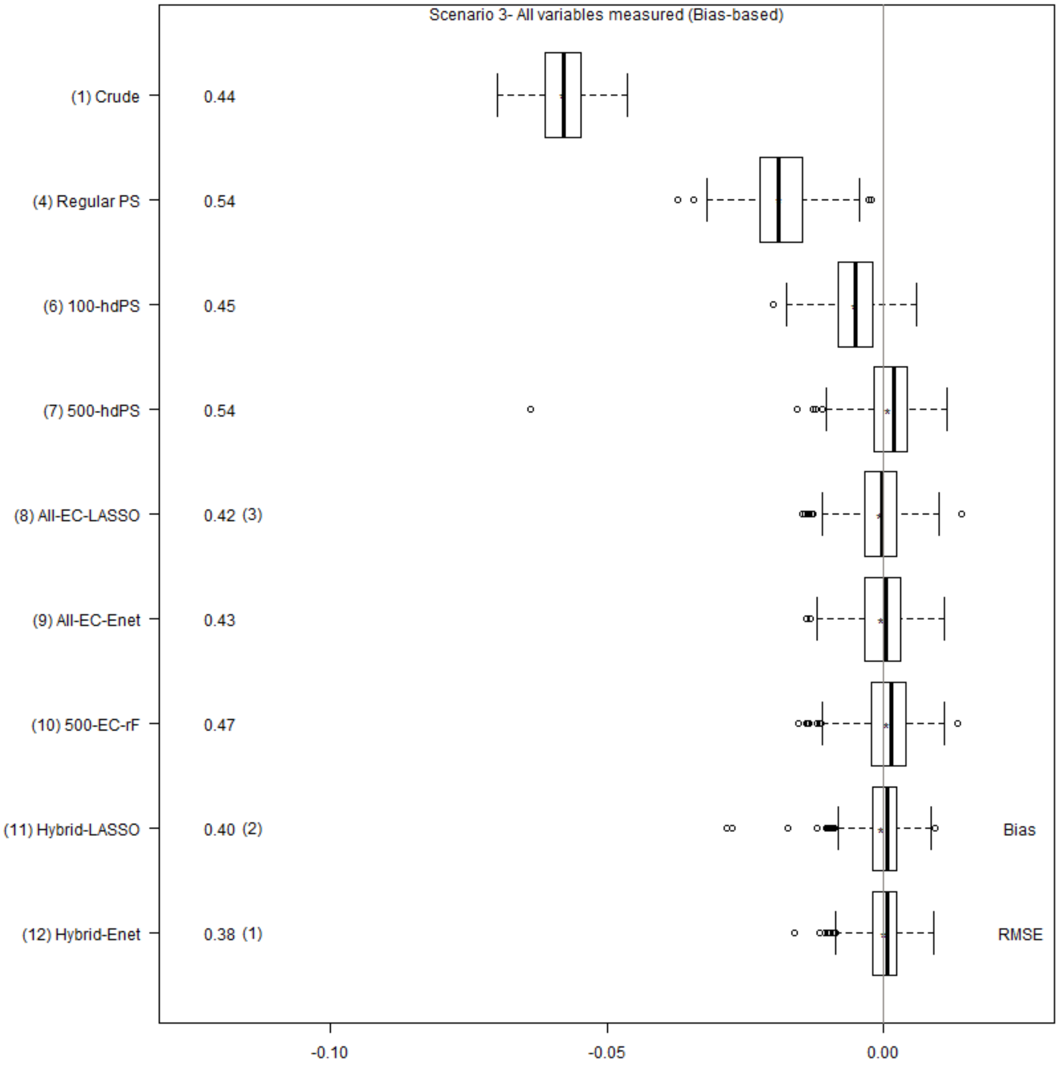
A.9.4 If all variables accounted (Bias-based analysis)



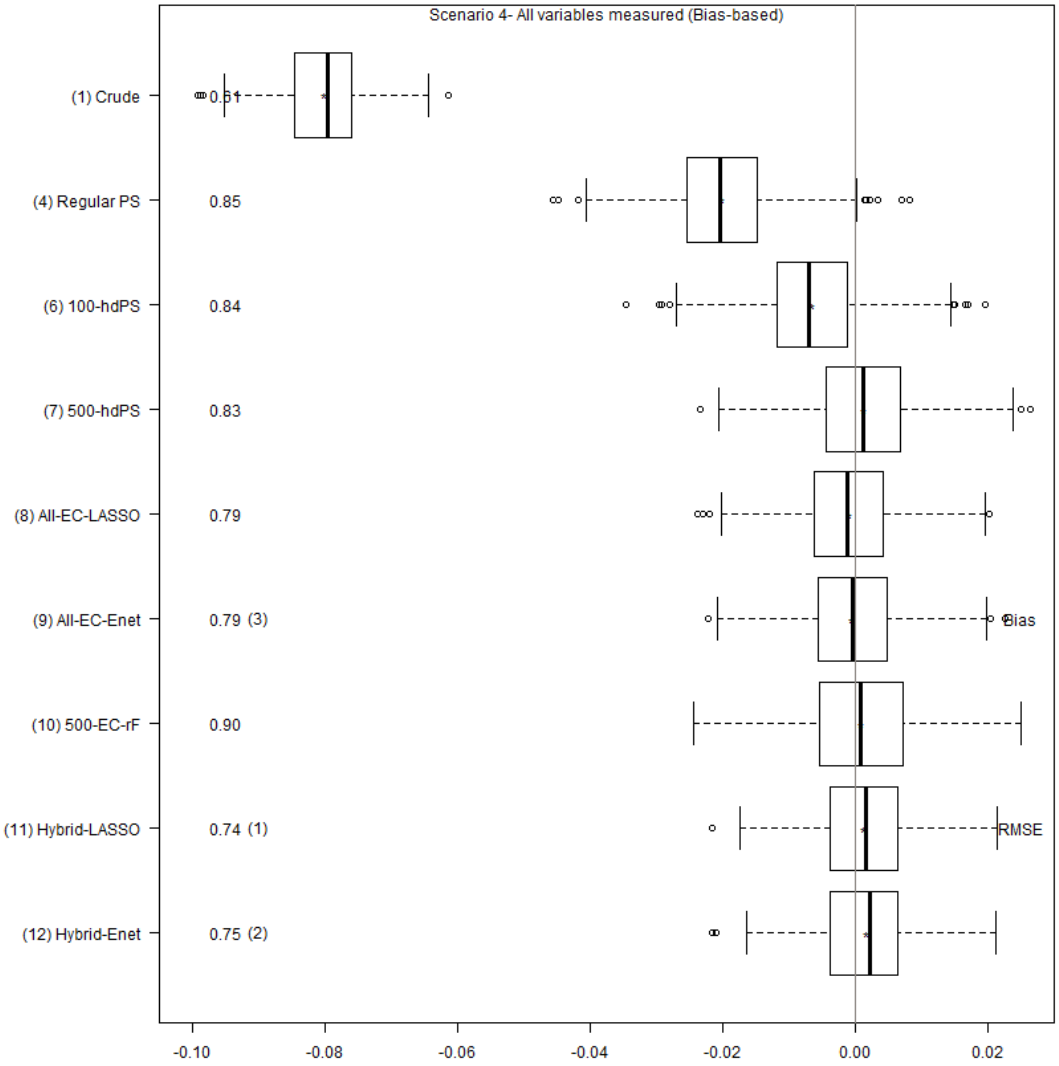
eFigure A.26: Plasmode Simulation Scenario 1-A



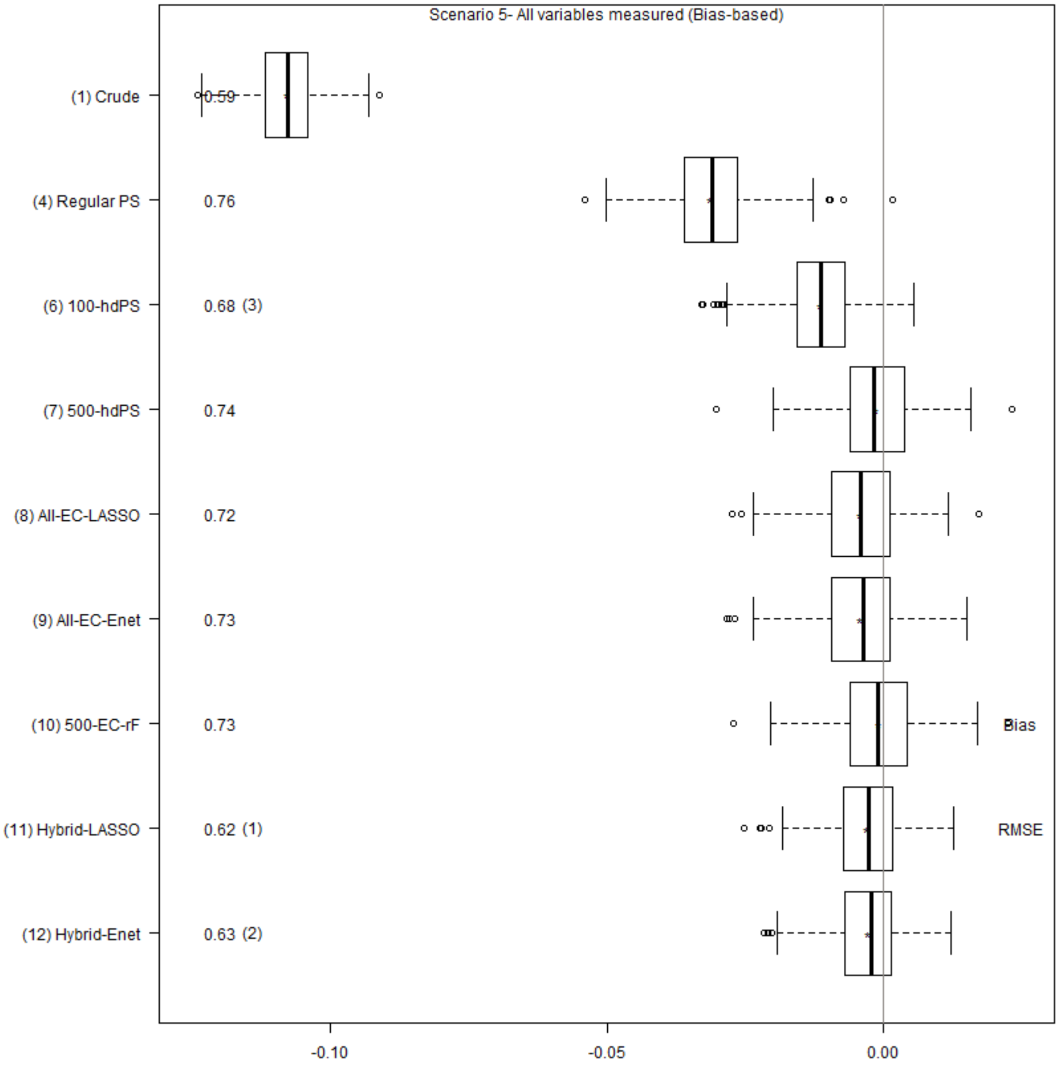
eFigure A.27: Plasmode Simulation Scenario 2-A



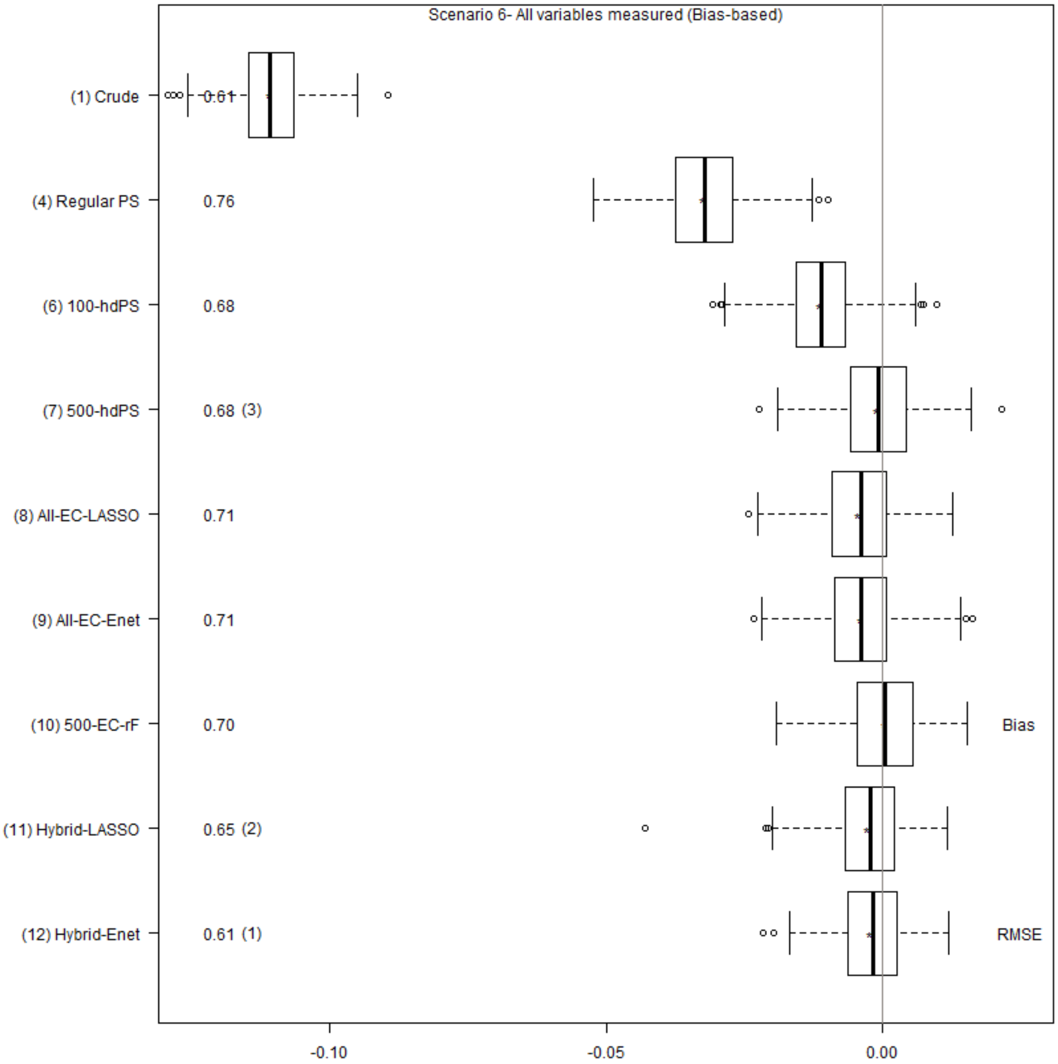
eFigure A.28: Plasmode Simulation Scenario 3-A



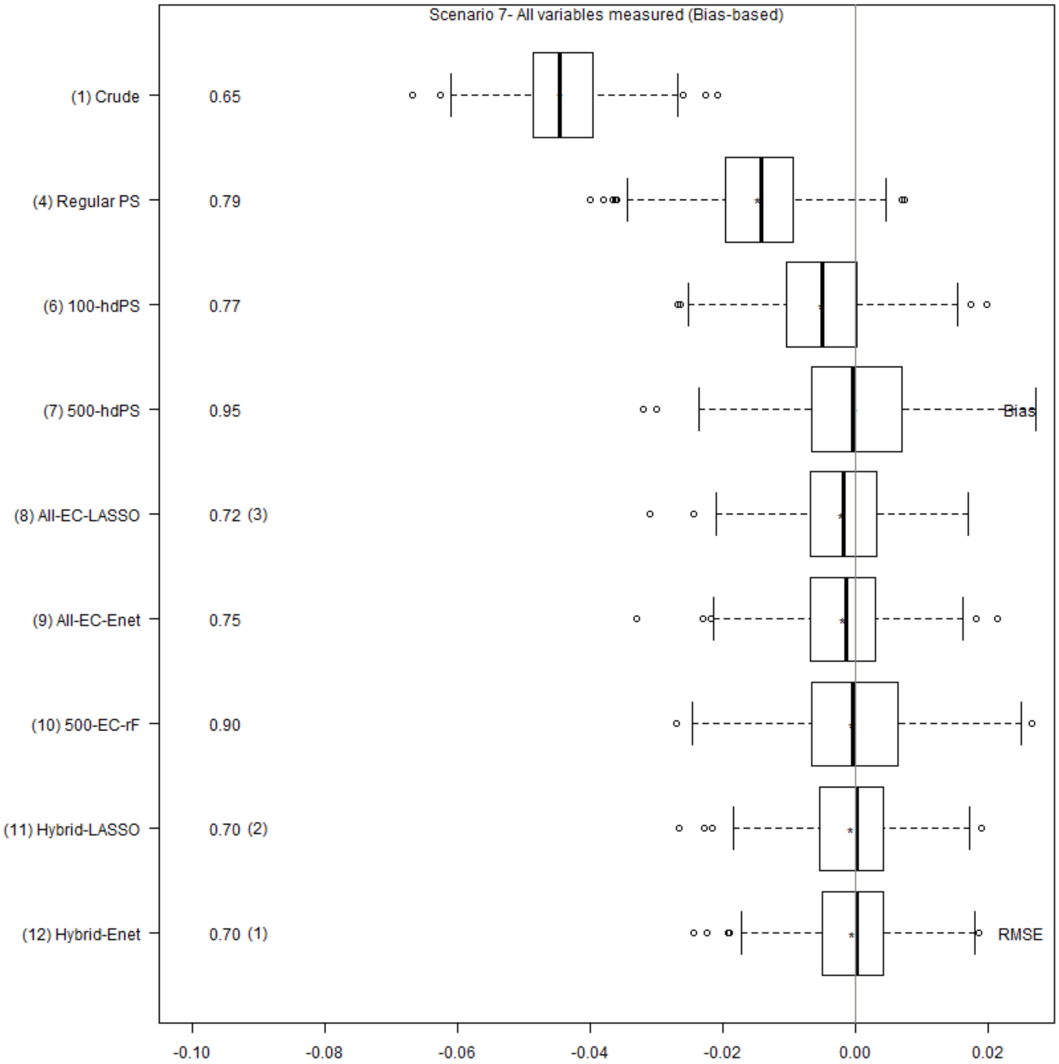
eFigure A.29: Plasmode Simulation Scenario 4-A



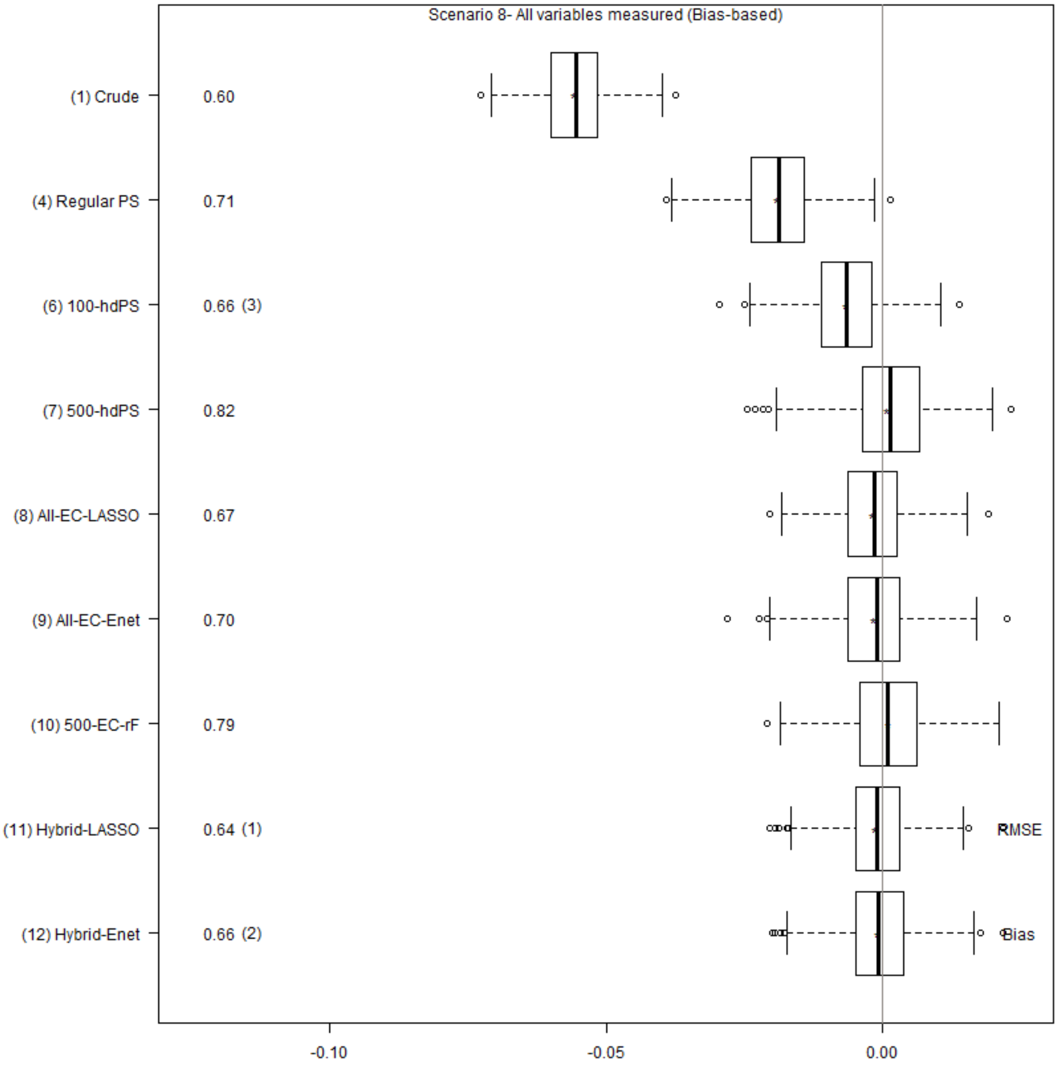
eFigure A.30: Plasmode Simulation Scenario 5-A



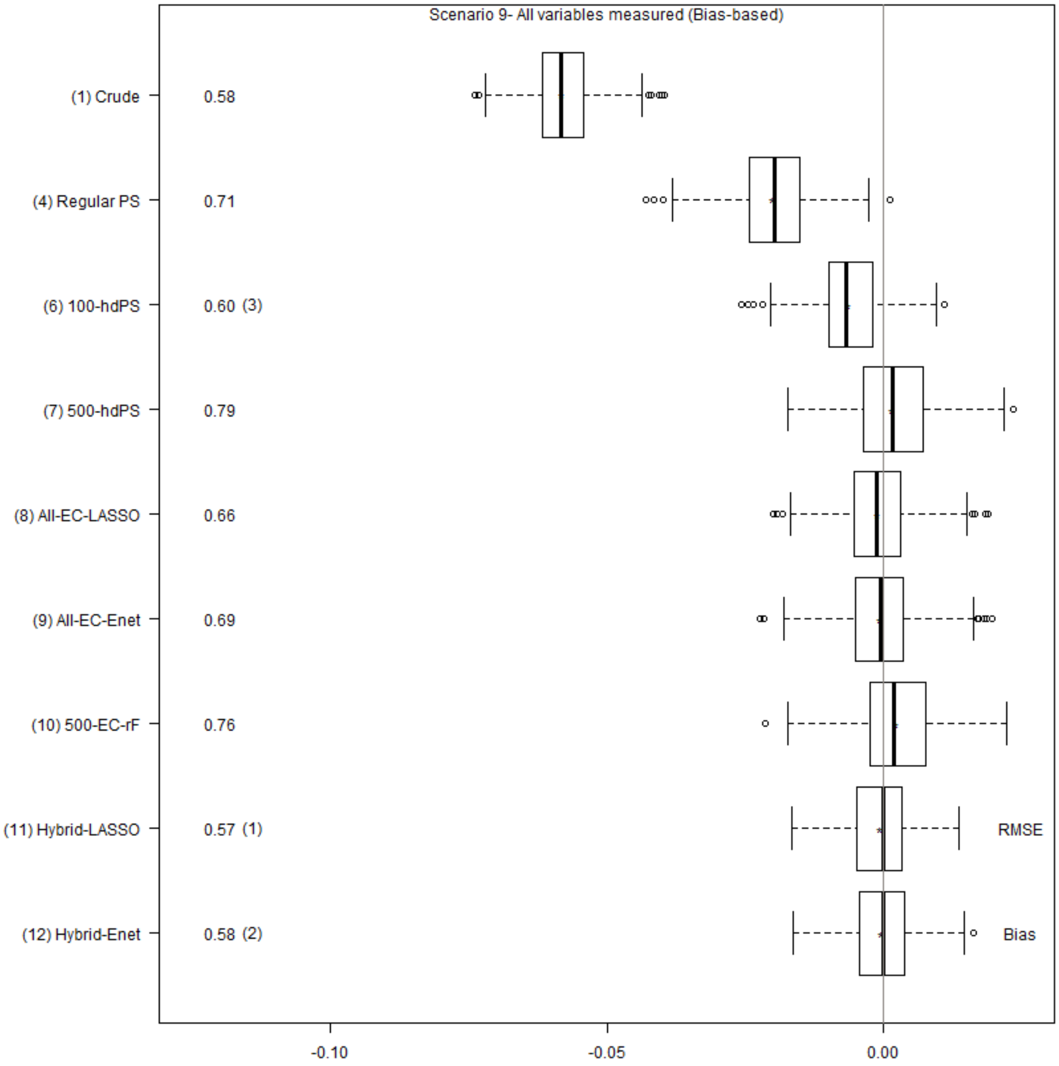
eFigure A.31: Plasmode Simulation Scenario 6-A



eFigure A.32: Plasmode Simulation Scenario 7-A

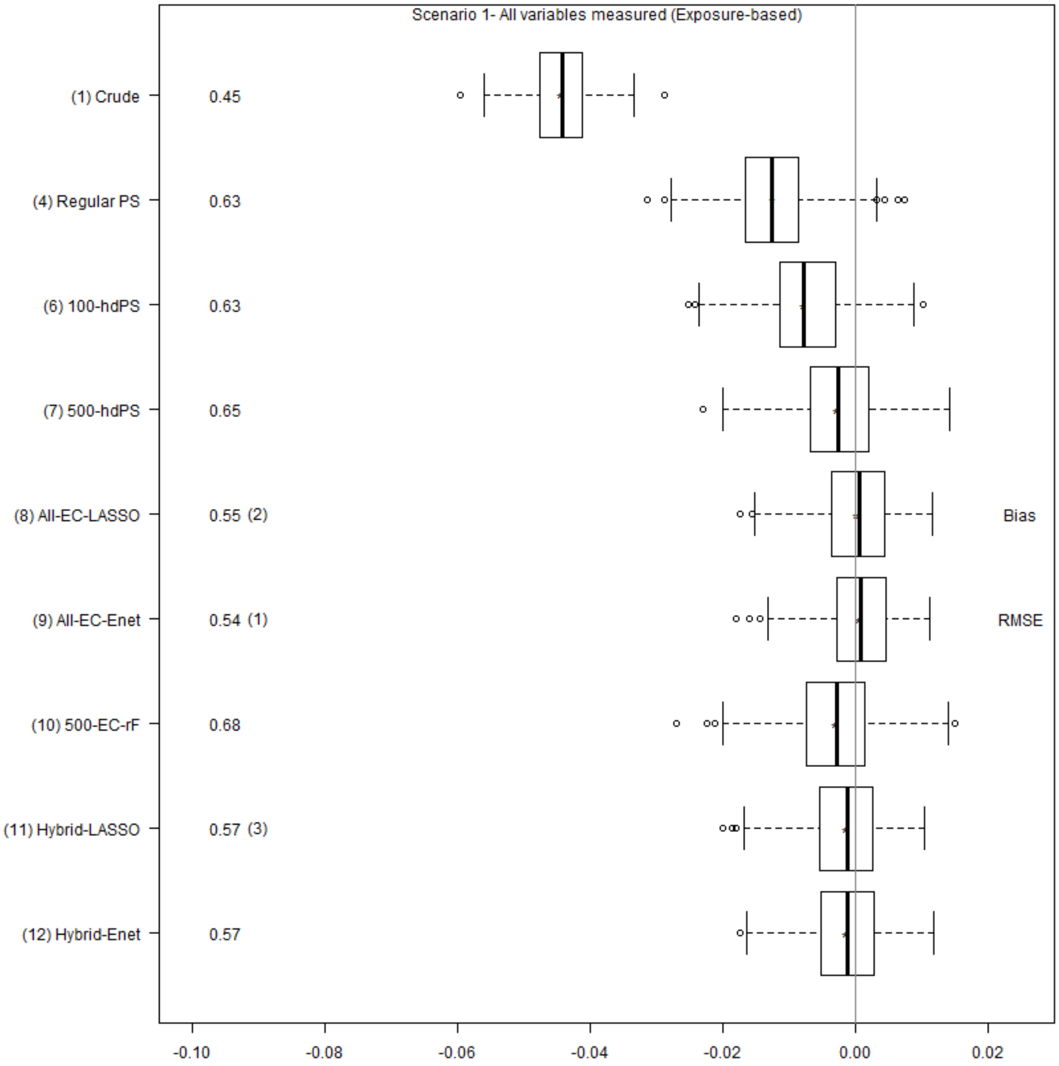


eFigure A.33: Plasmode Simulation Scenario 8-A

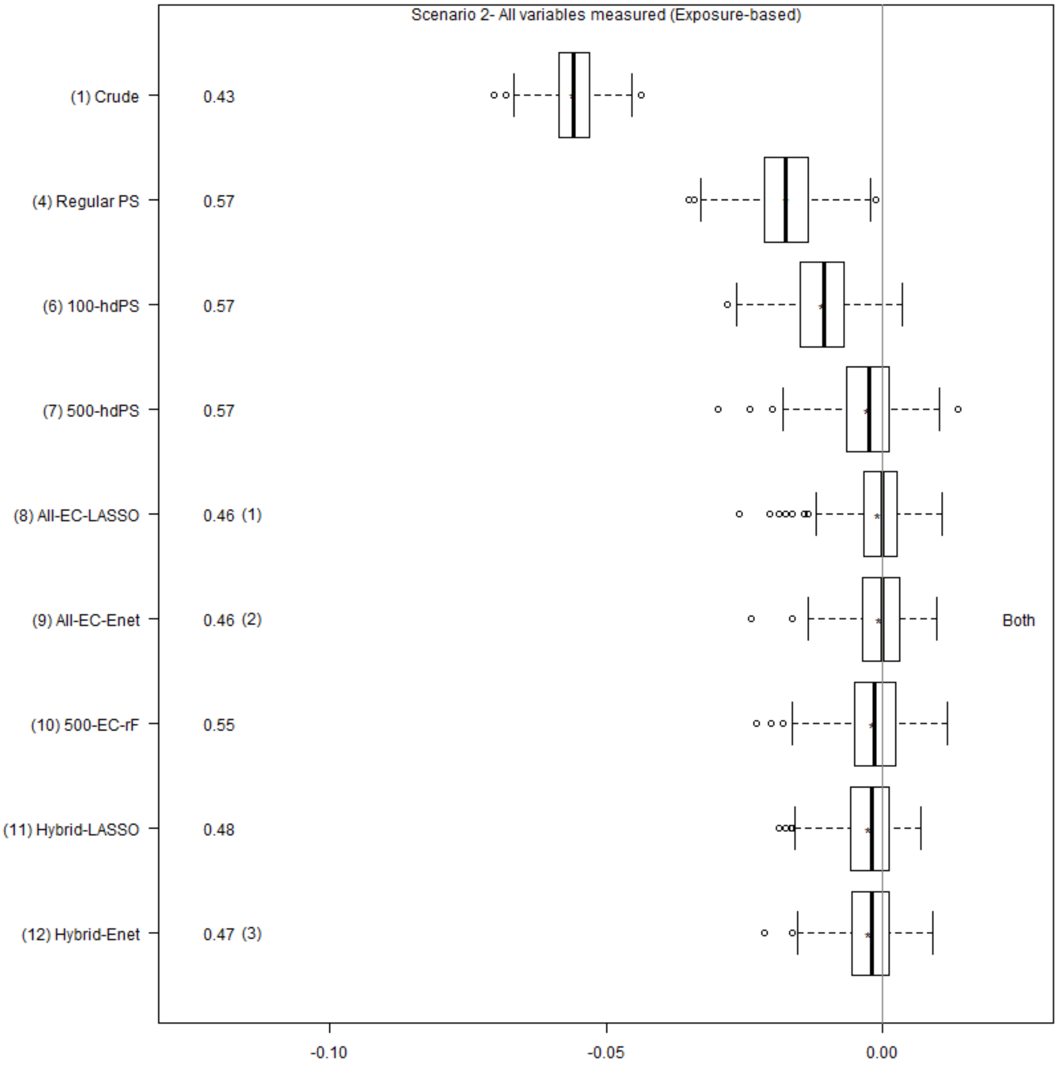


eFigure A.34: Plasmode Simulation Scenario 9-A

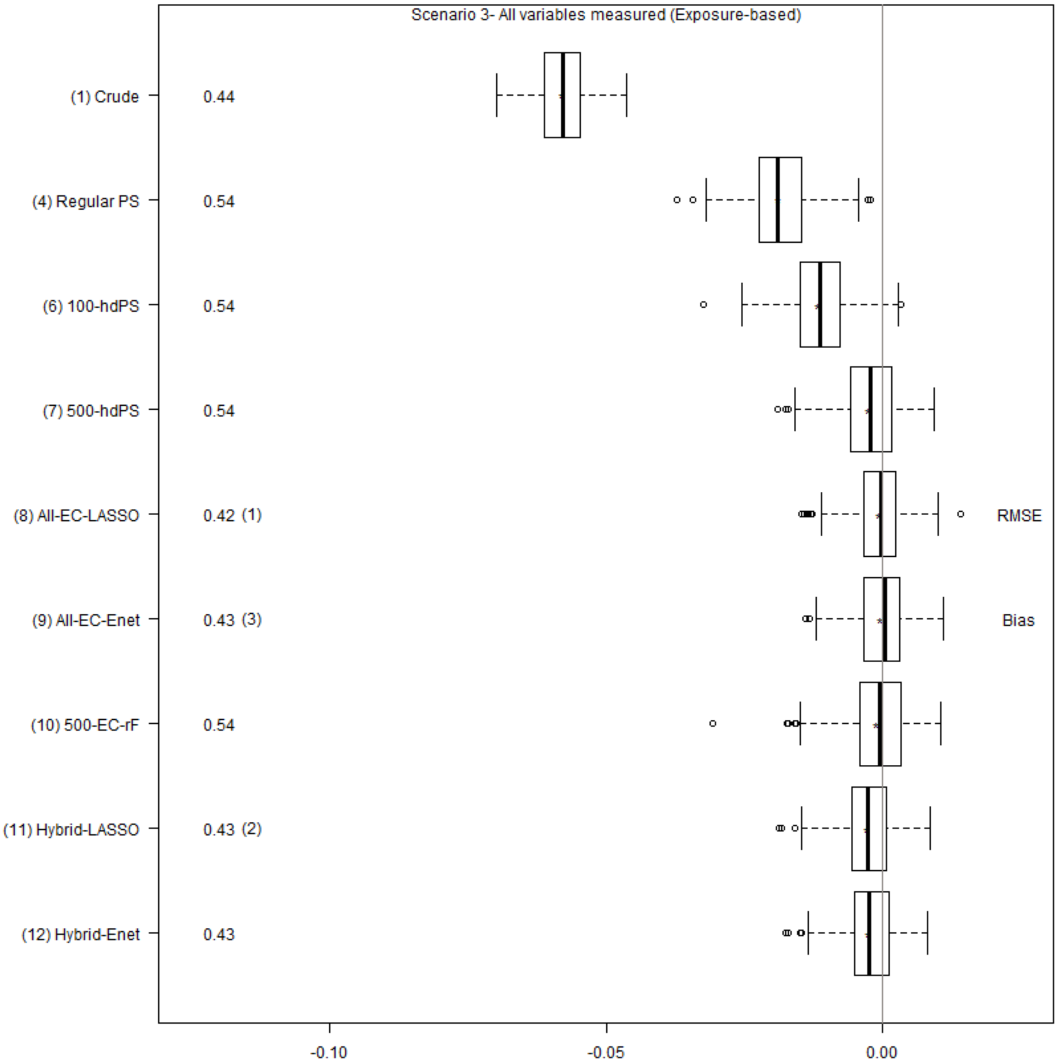
A.9.5 If all variables accounted (Exposure-based analysis)



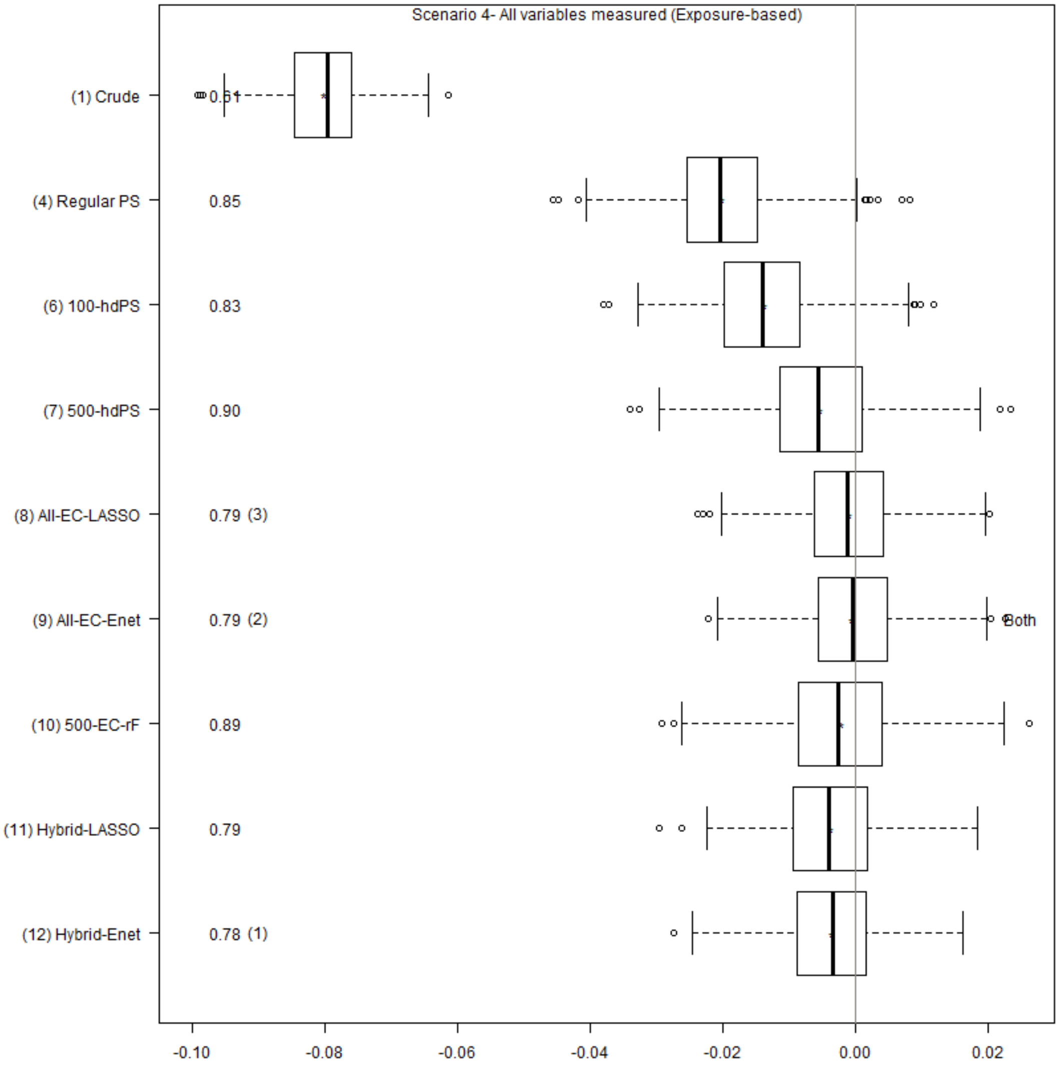
eFigure A.35: Plasmode Simulation Scenario 1-A



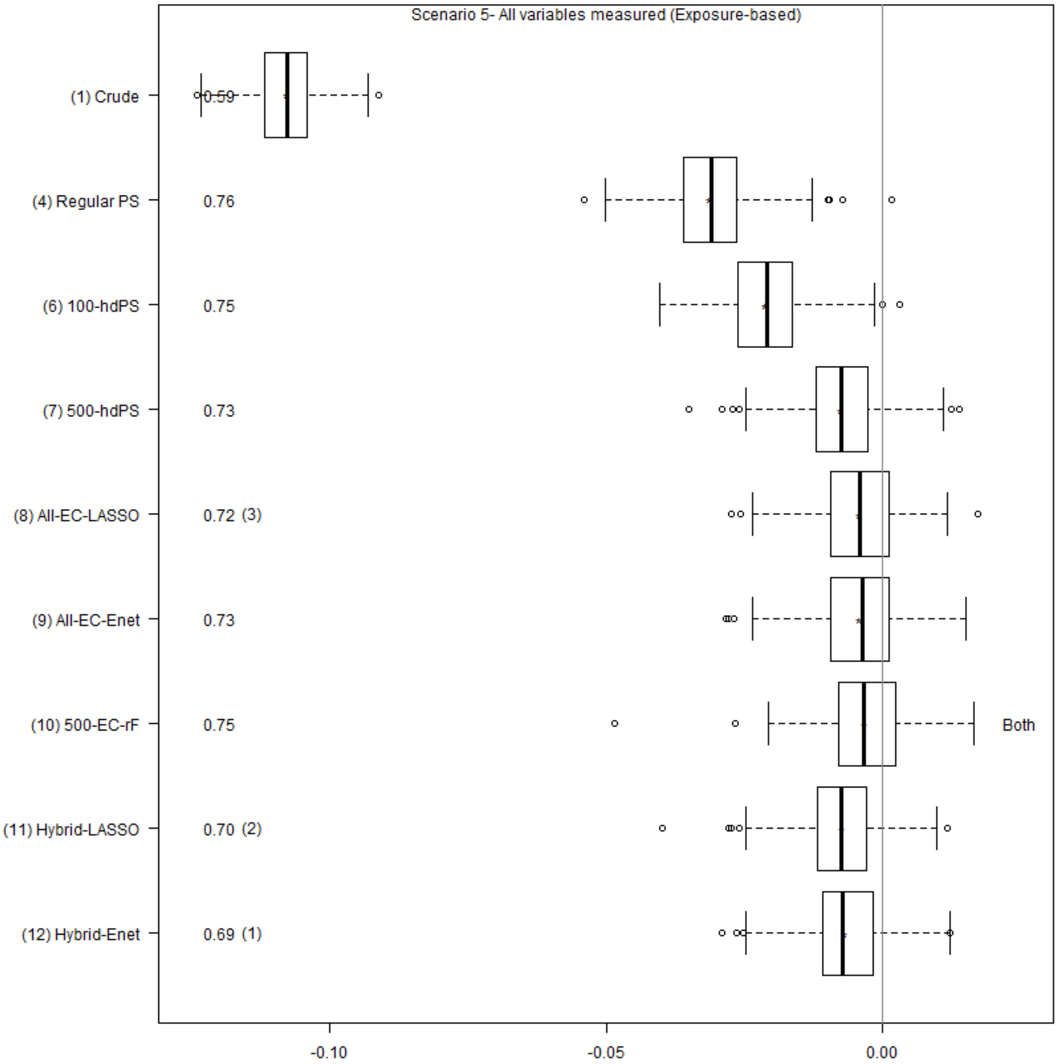
eFigure A.36: Plasmode Simulation Scenario 2-A



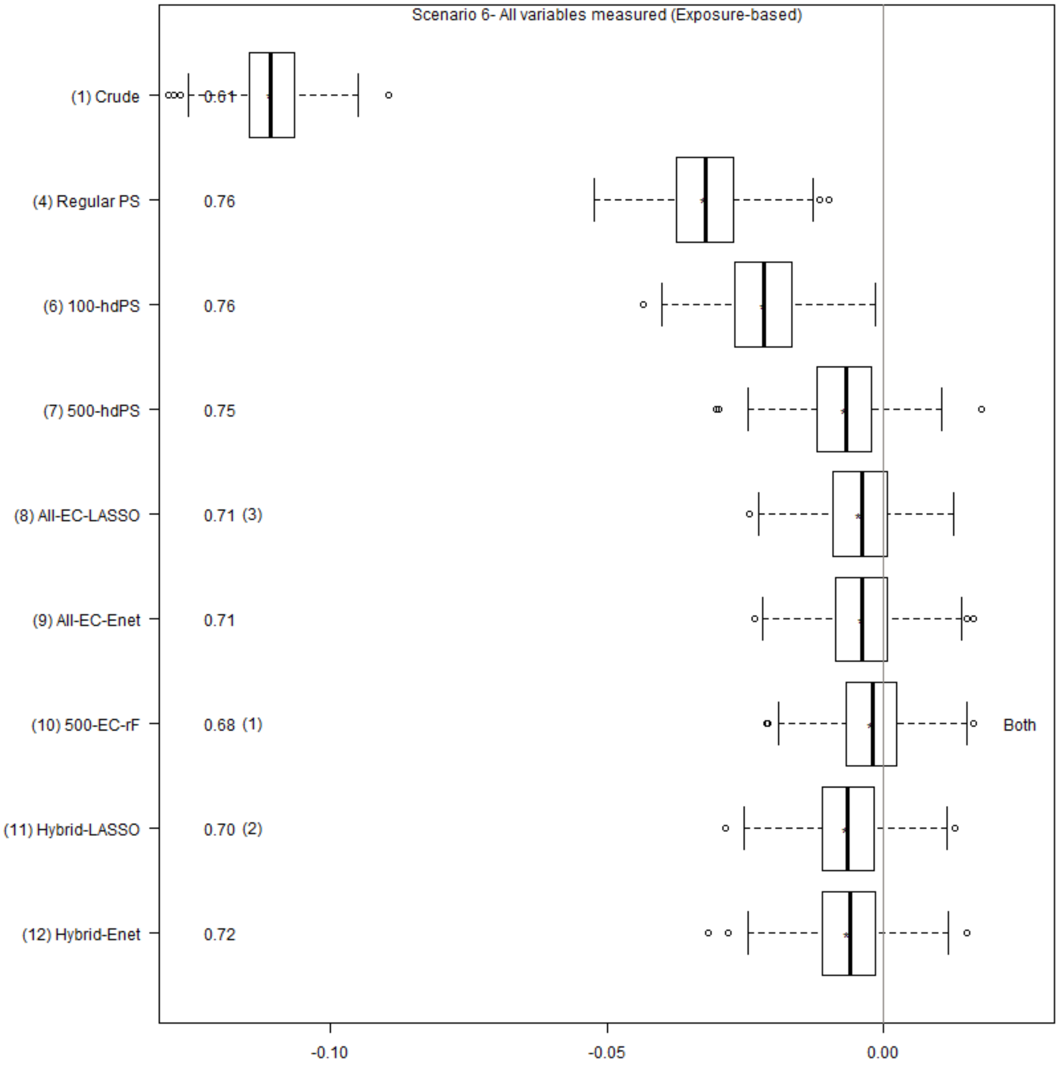
eFigure A.37: Plasmode Simulation Scenario 3-A



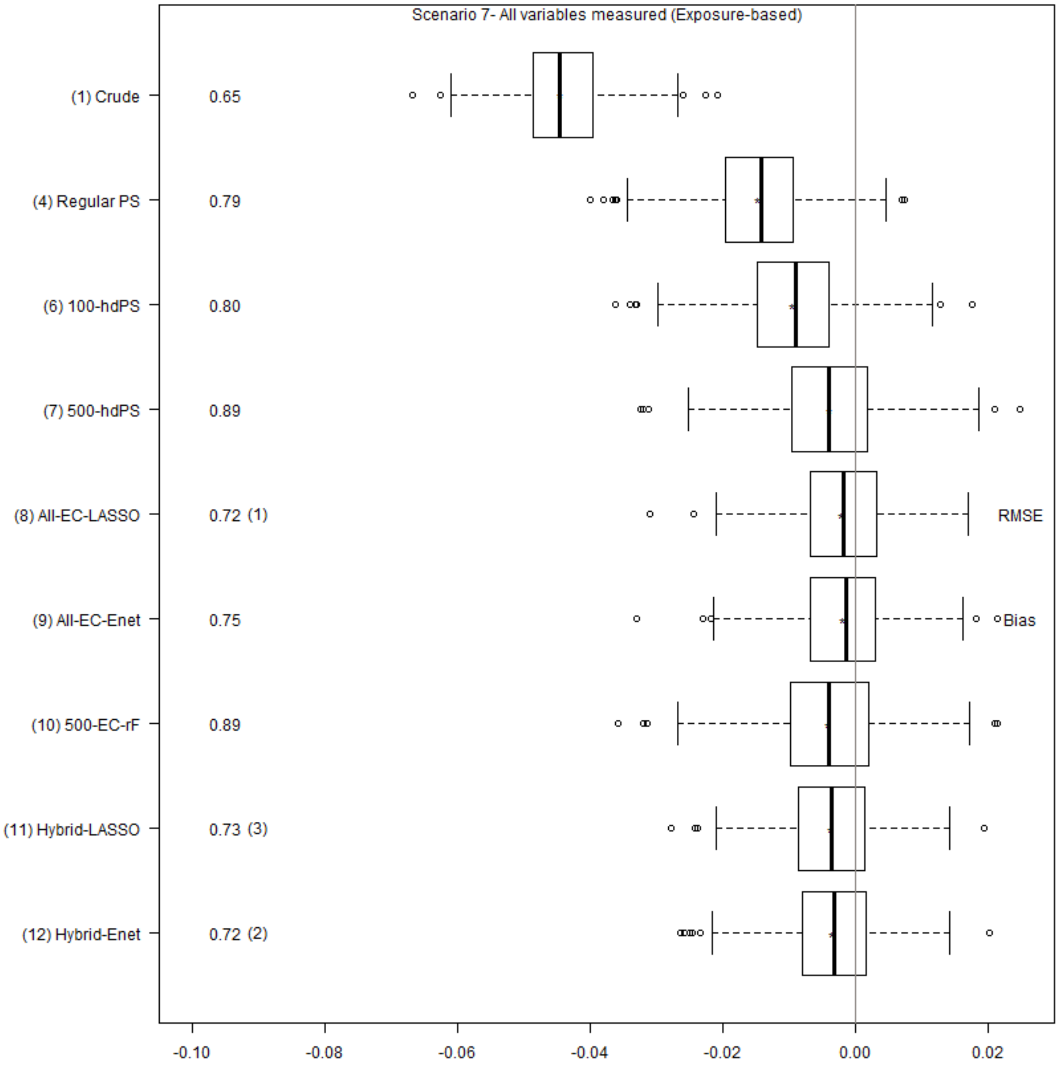
eFigure A.38: Plasmode Simulation Scenario 4-A



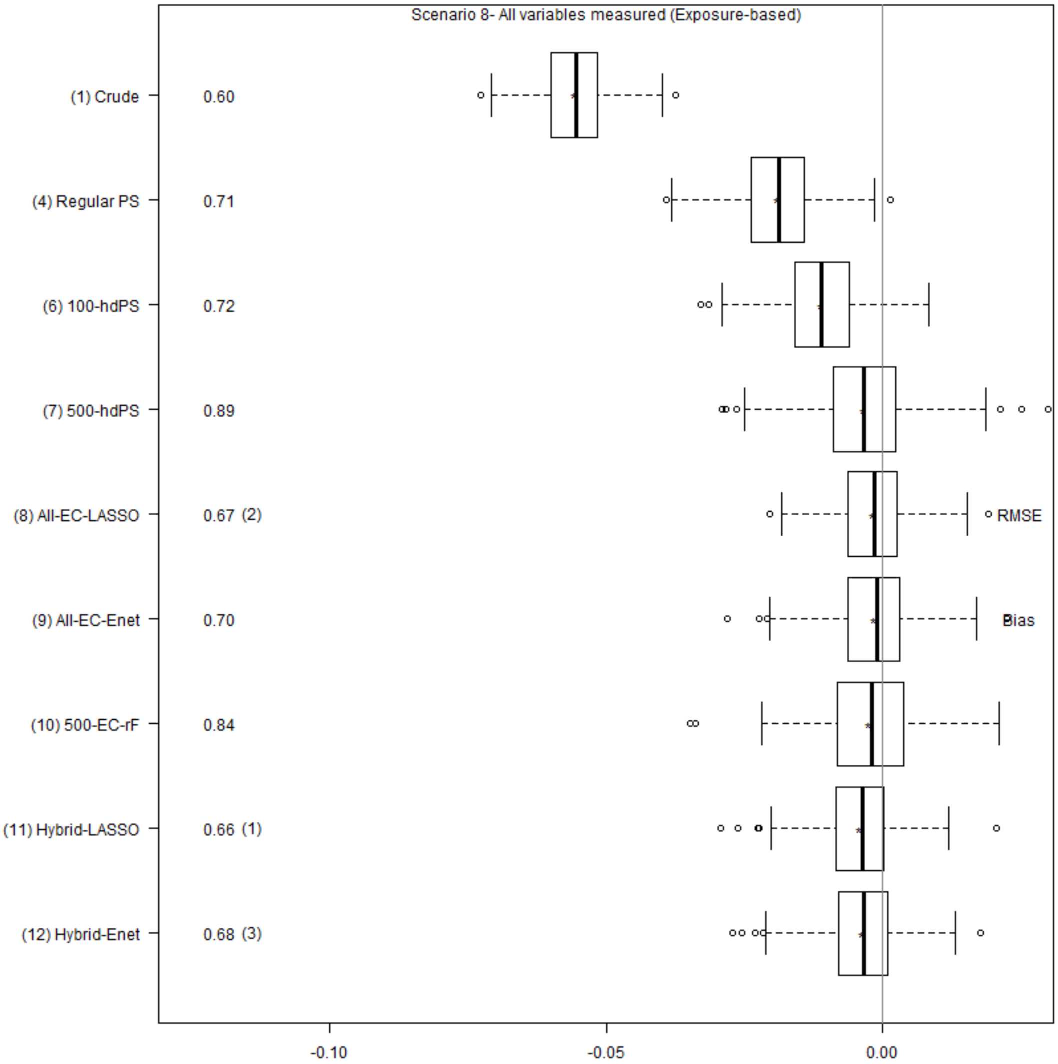
eFigure A.39: Plasmode Simulation Scenario 5-A



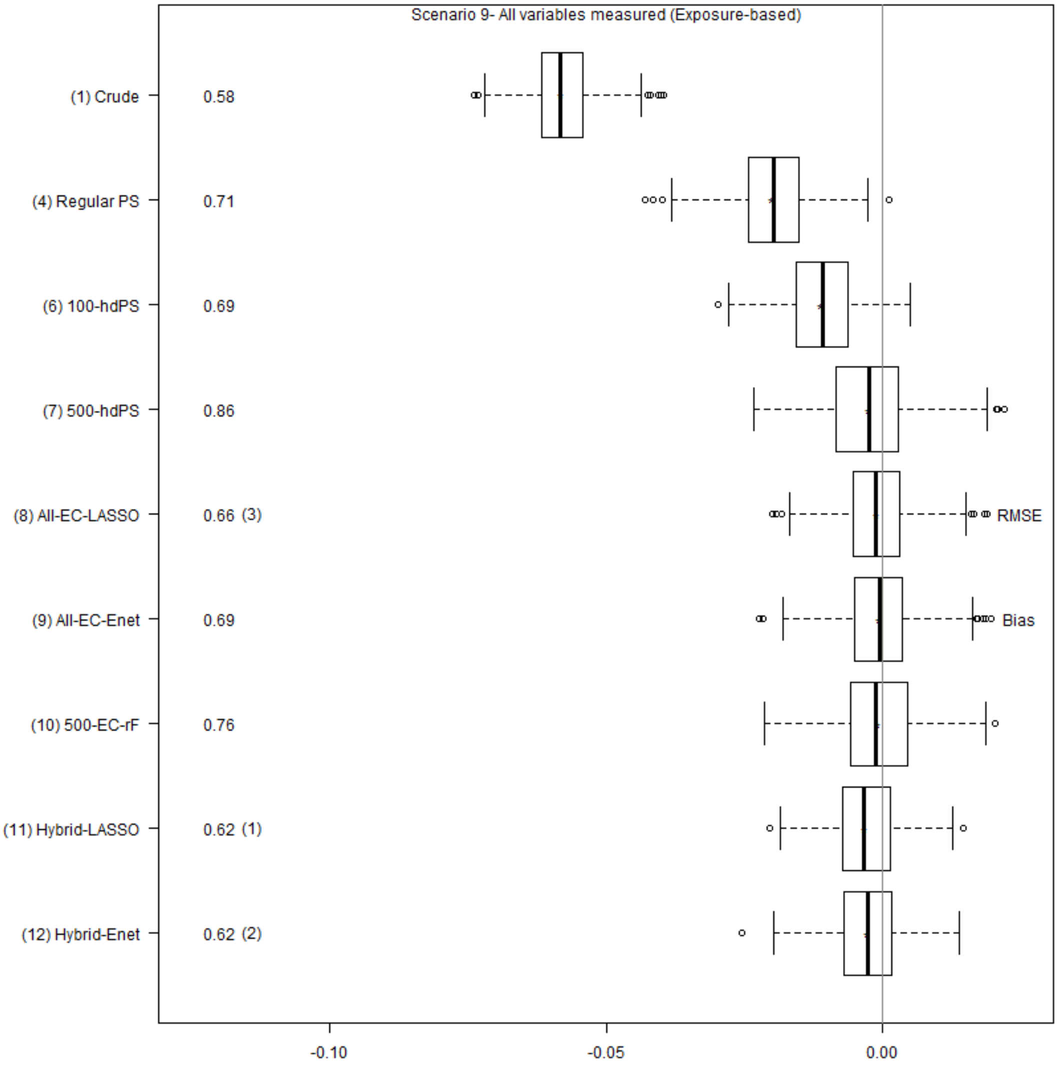
eFigure A.40: Plasmode Simulation Scenario 6-A



eFigure A.41: Plasmode Simulation Scenario 7-A



eFigure A.42: Plasmode Simulation Scenario 8-A

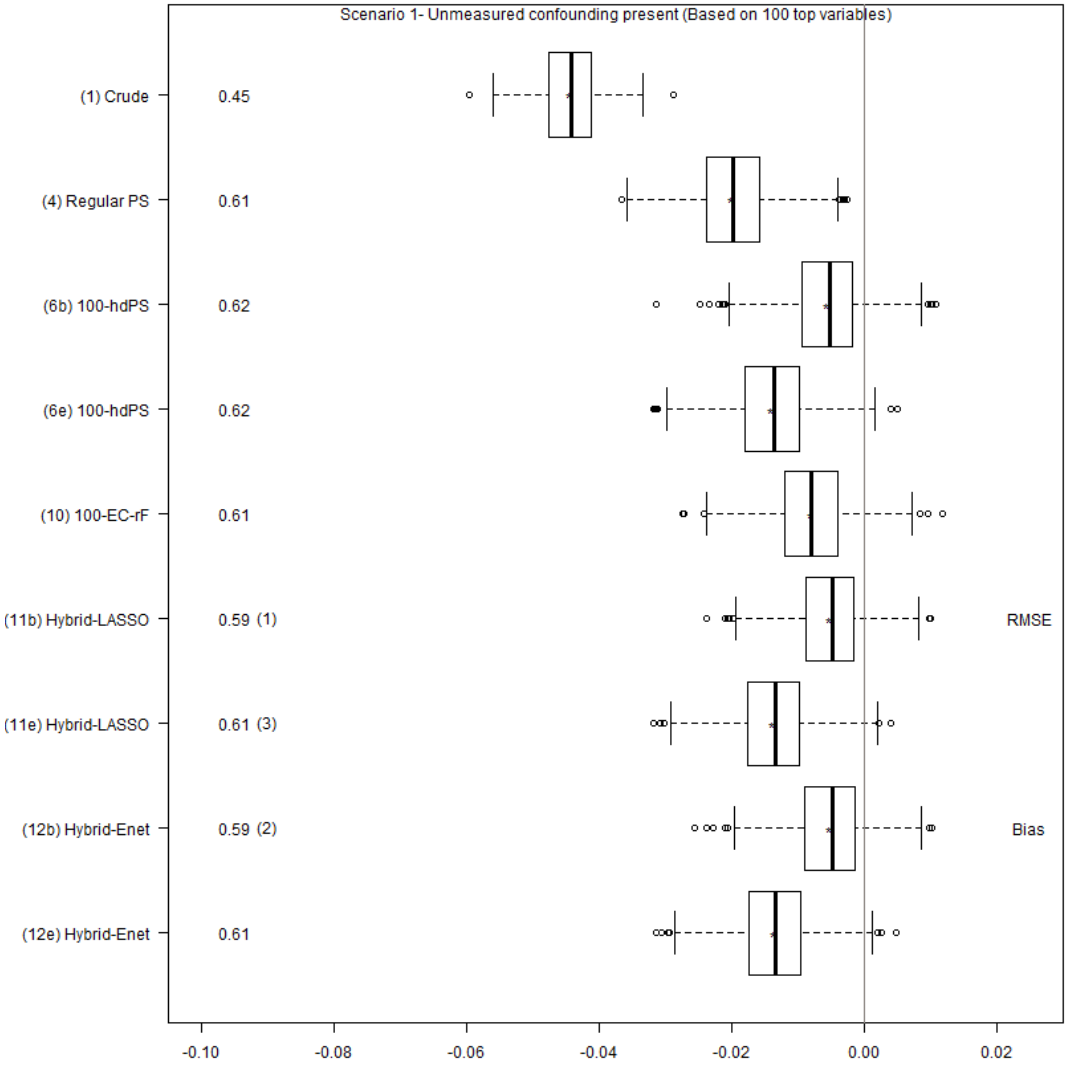


eFigure A.43: Plasmode Simulation Scenario 9-A

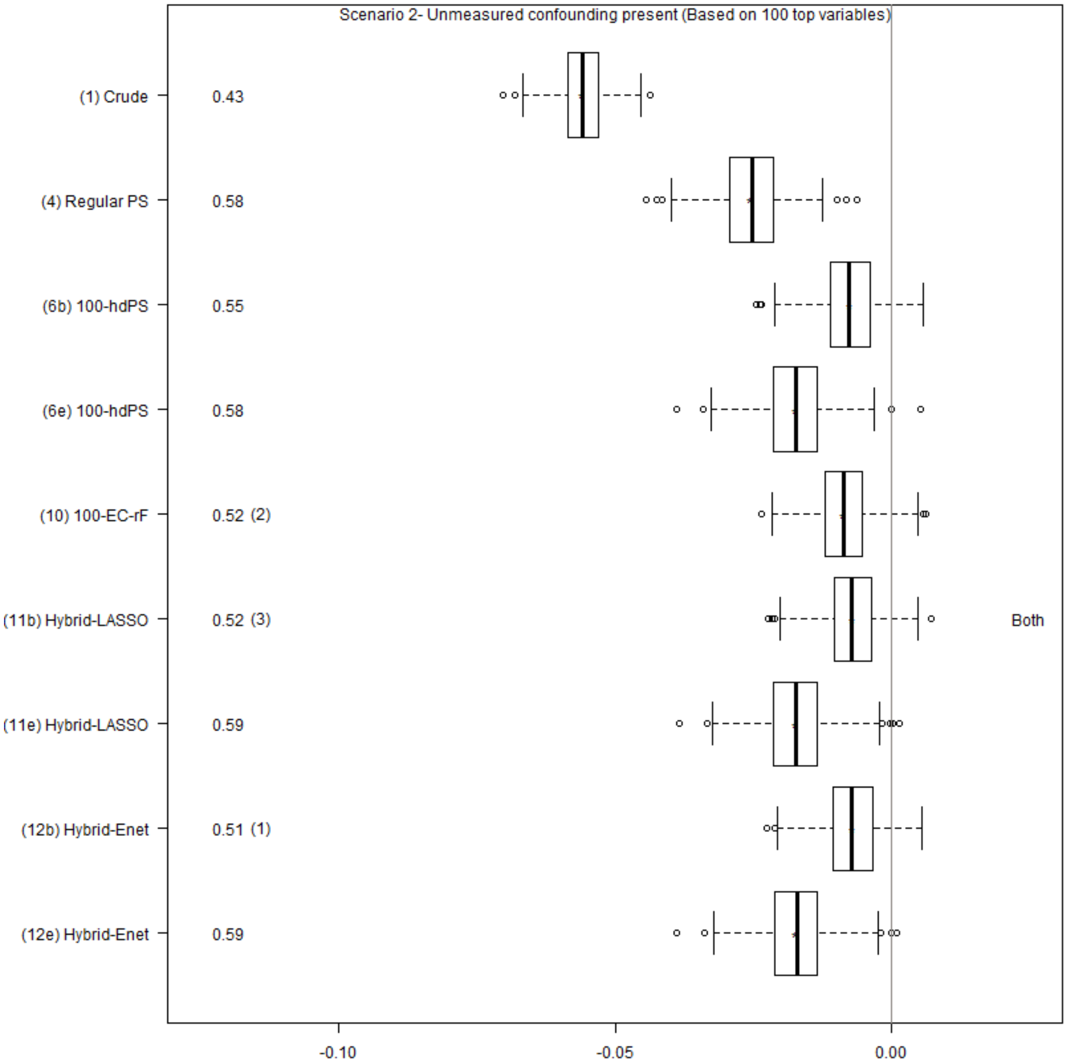
A.9.6 Considering fewer variables in the analysis

When the same simulated scenarios were analyzed based on only 100 top hdPS variables, generally, more bias is associated in the treatment effect estimation, but hybrid methods (Hybrid-Enet and Hybrid-LASSO based on 100 hdPS variables) continue to dominate almost all the scenarios (see eFigures A.44-A.52 and eFigures A.53-A.61). Only in a few cases with amplified confounding effect ($\gamma = 3$ or 5), 100-EC-rF performed best when the analysis was based on exposure-based ranking and in two cases, 100-hdPS performed best when bias-based ranking was conducted.

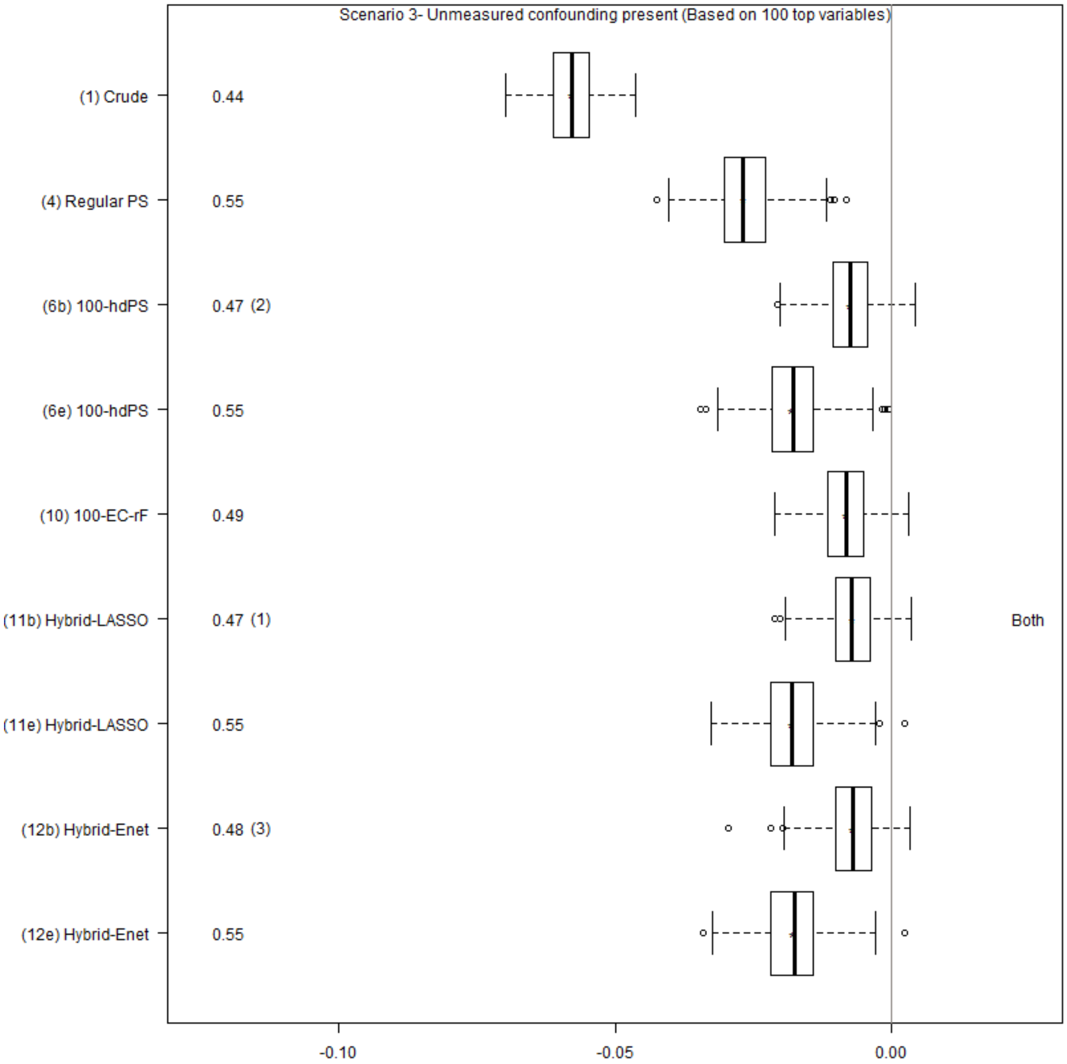
A.9.7 If unmeasured confounding present (Based on top 100 selected variables)



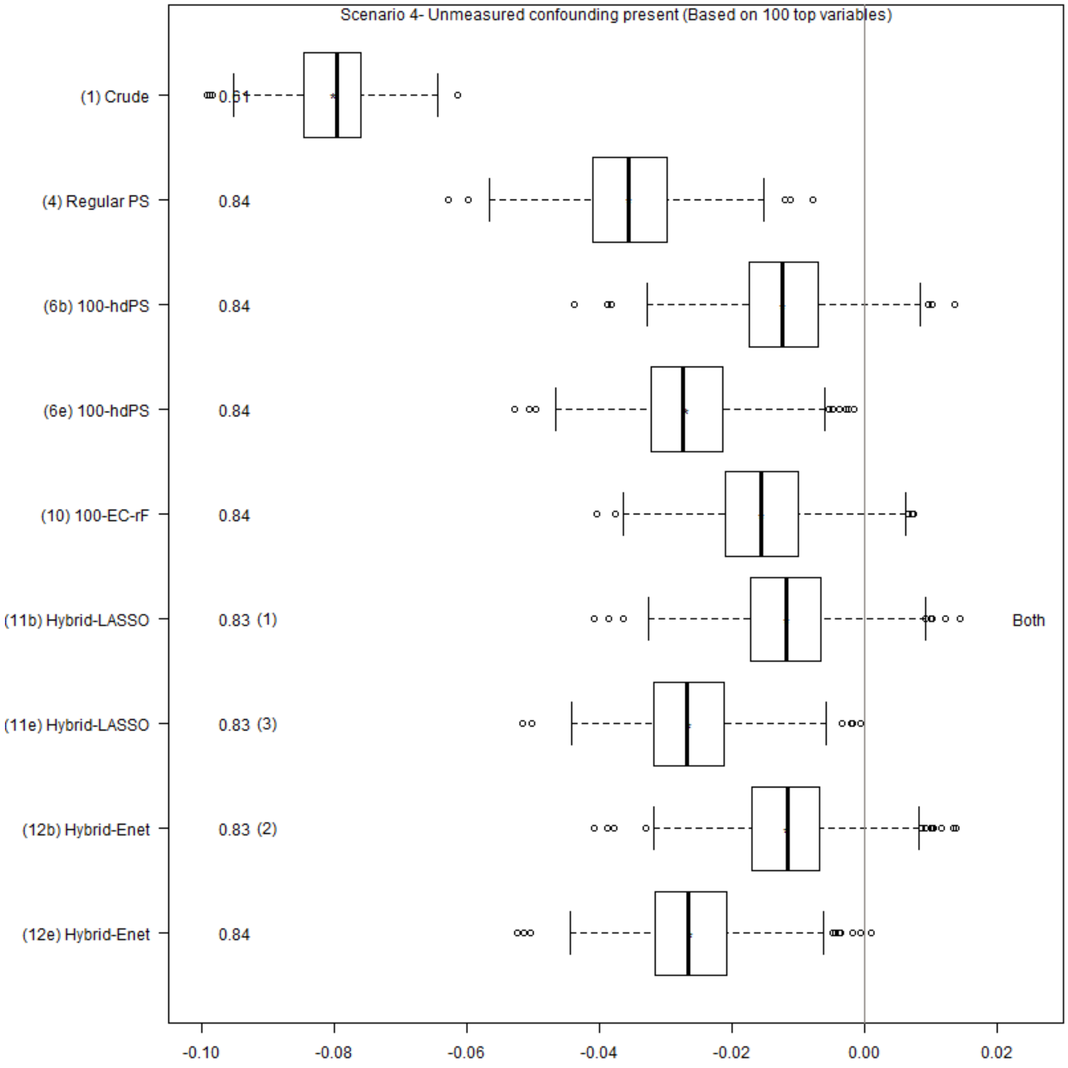
eFigure A.44: Plasmode Simulation Scenario 1-A



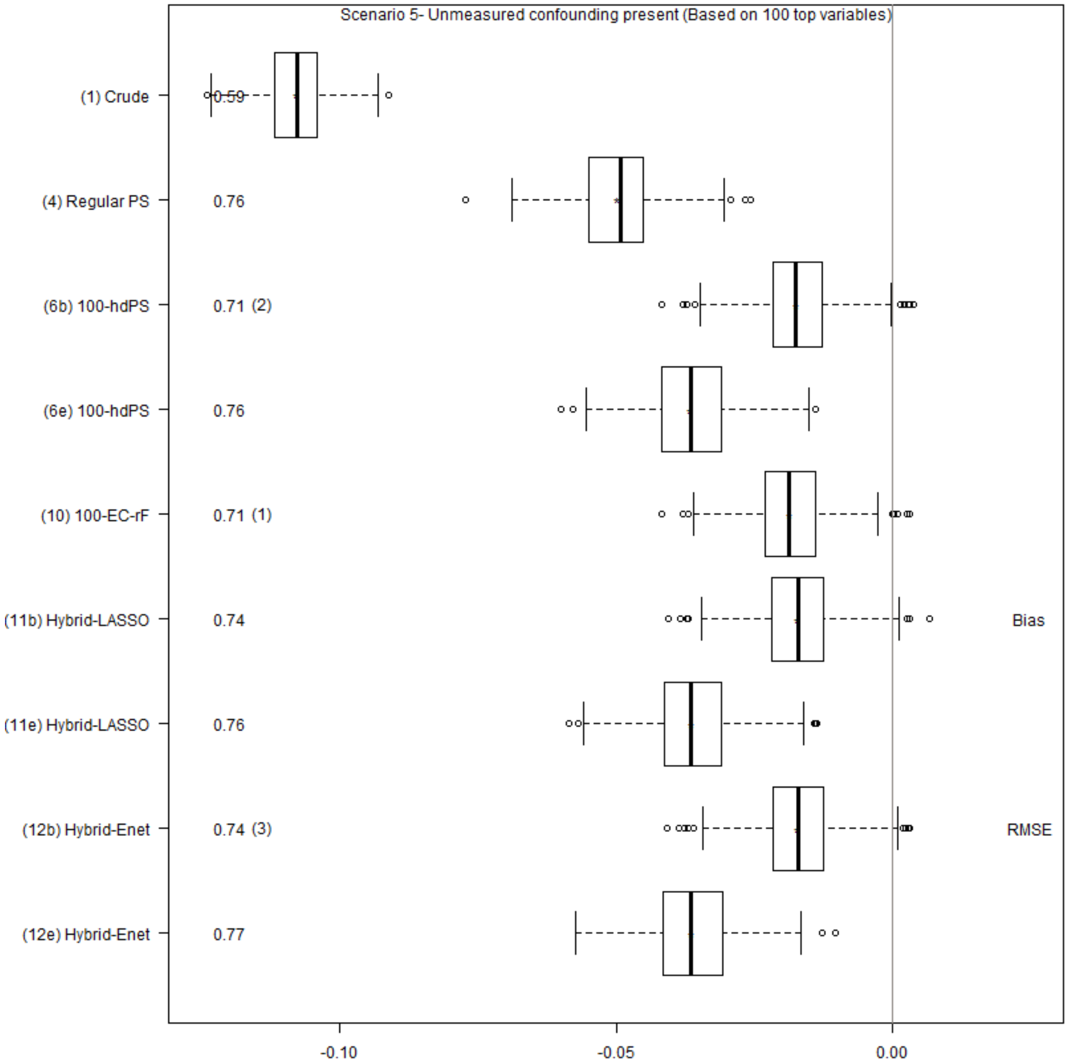
eFigure A.45: Plasmode Simulation Scenario 2-A



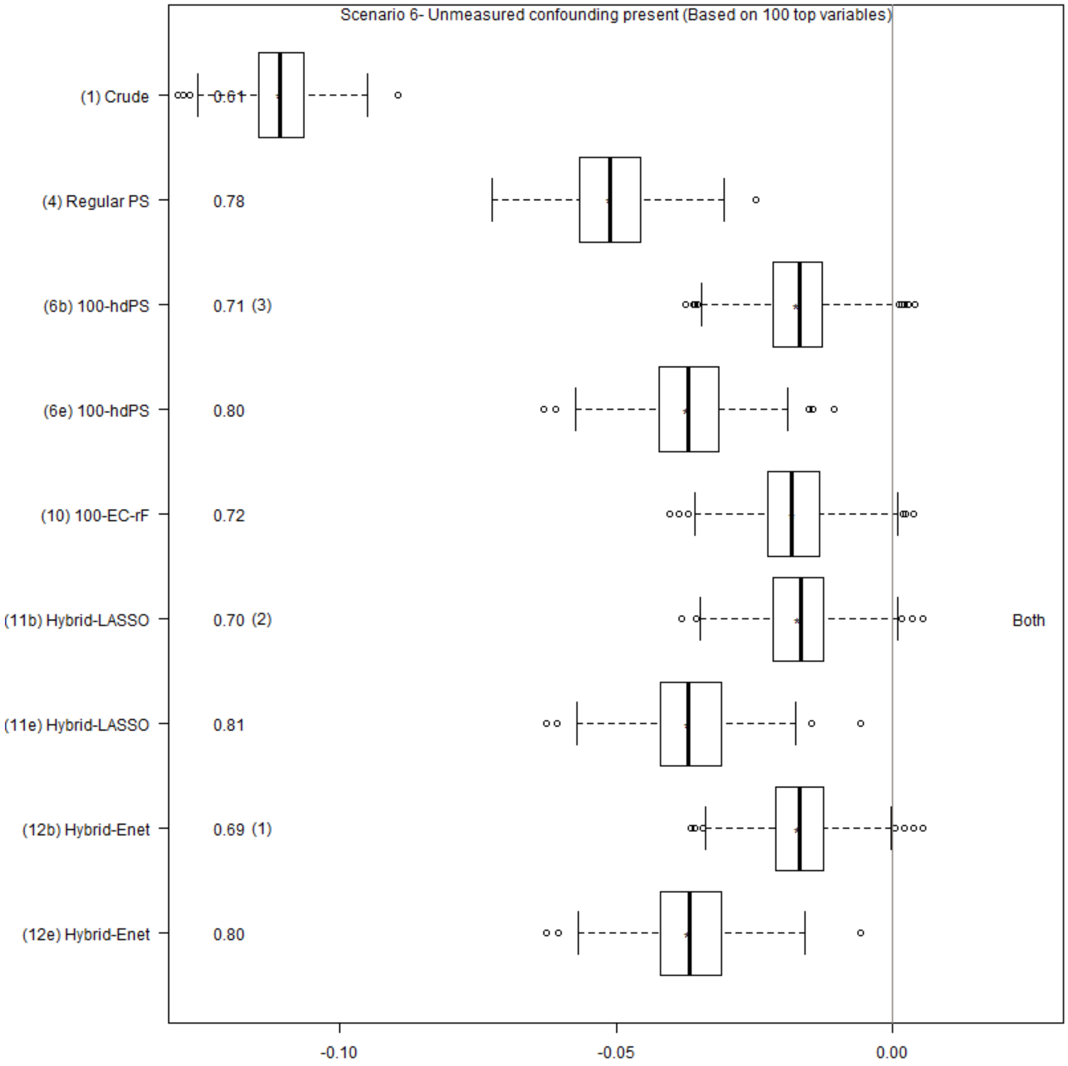
eFigure A.46: Plasmode Simulation Scenario 3-A



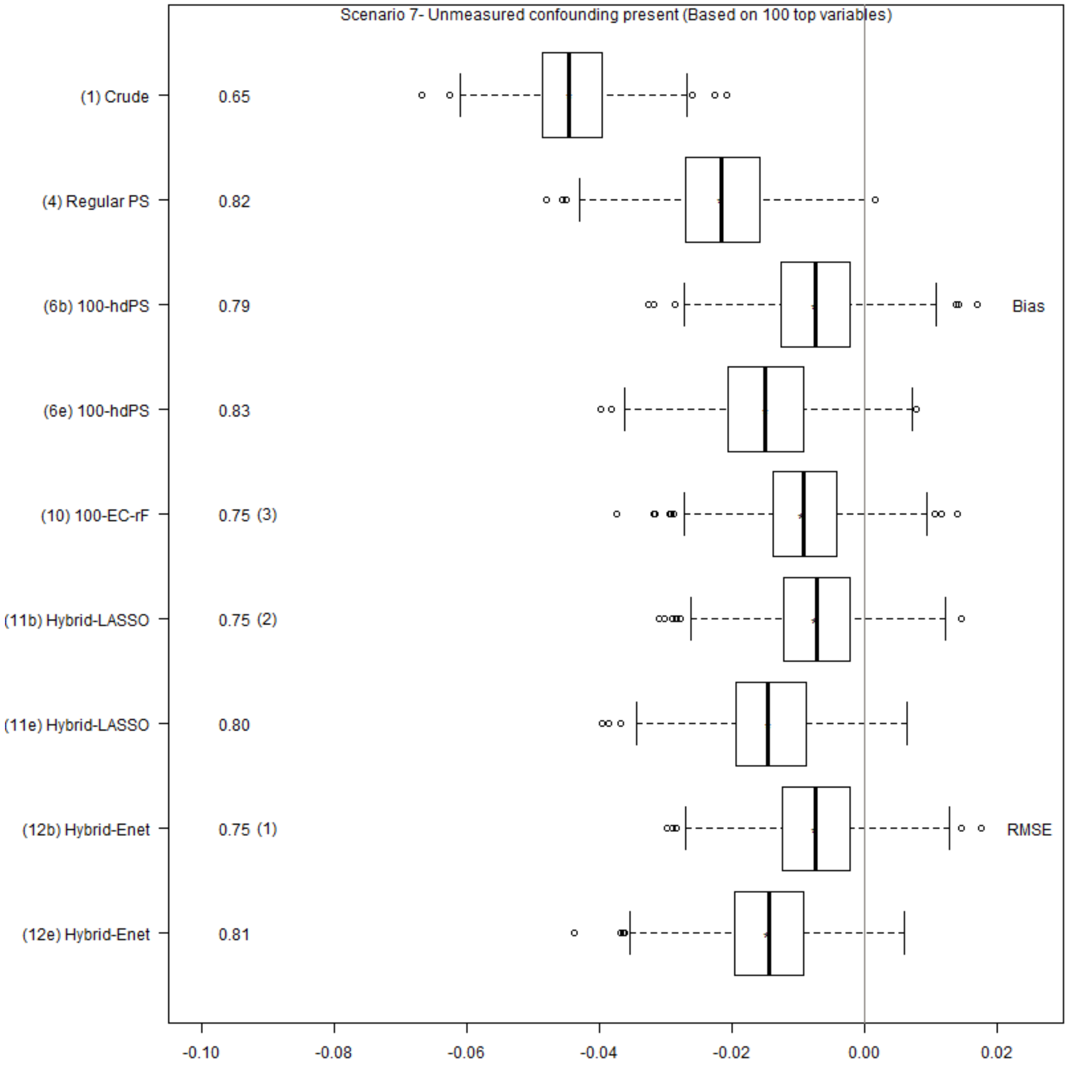
eFigure A.47: Plasmode Simulation Scenario 4-A



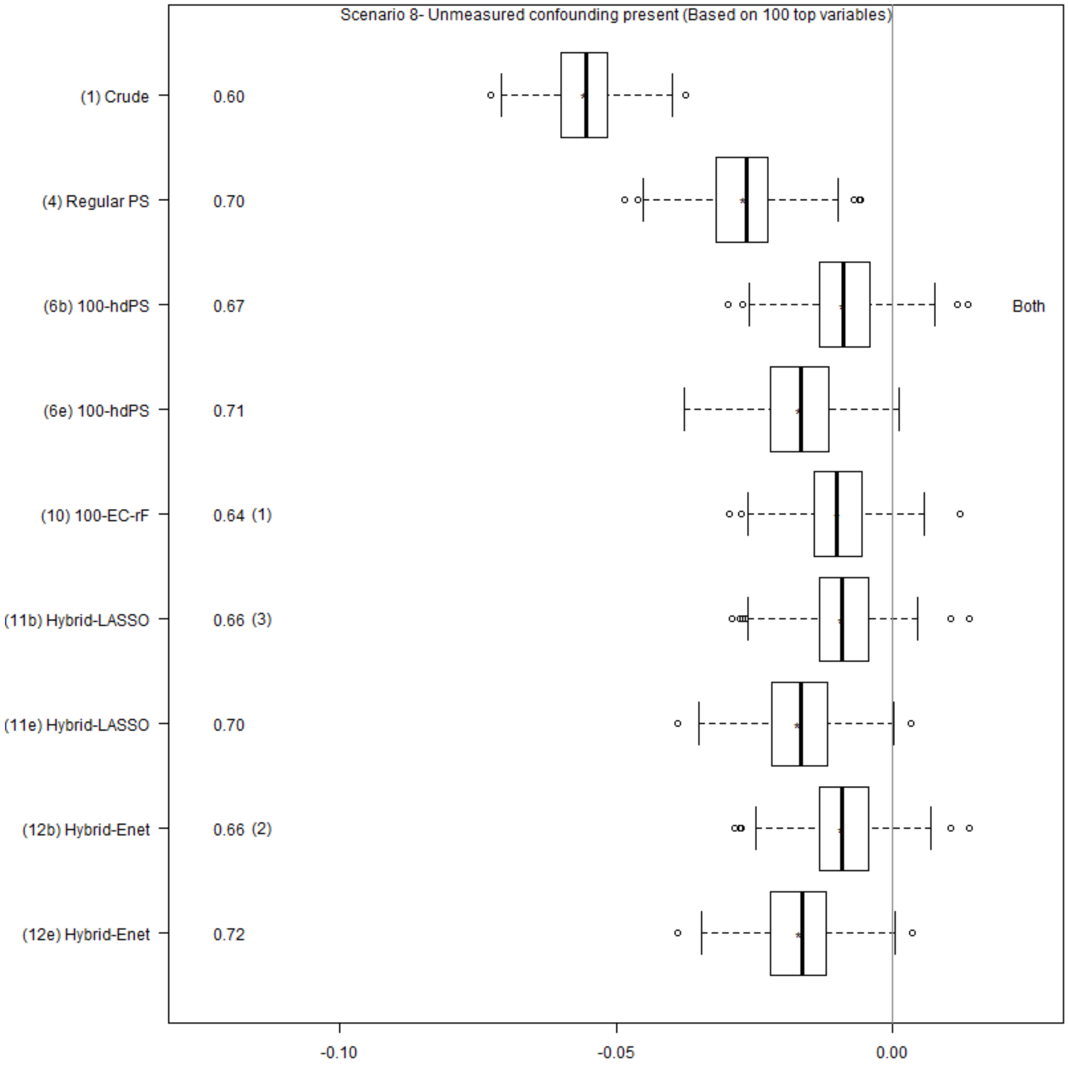
eFigure A.48: Plasmode Simulation Scenario 5-A



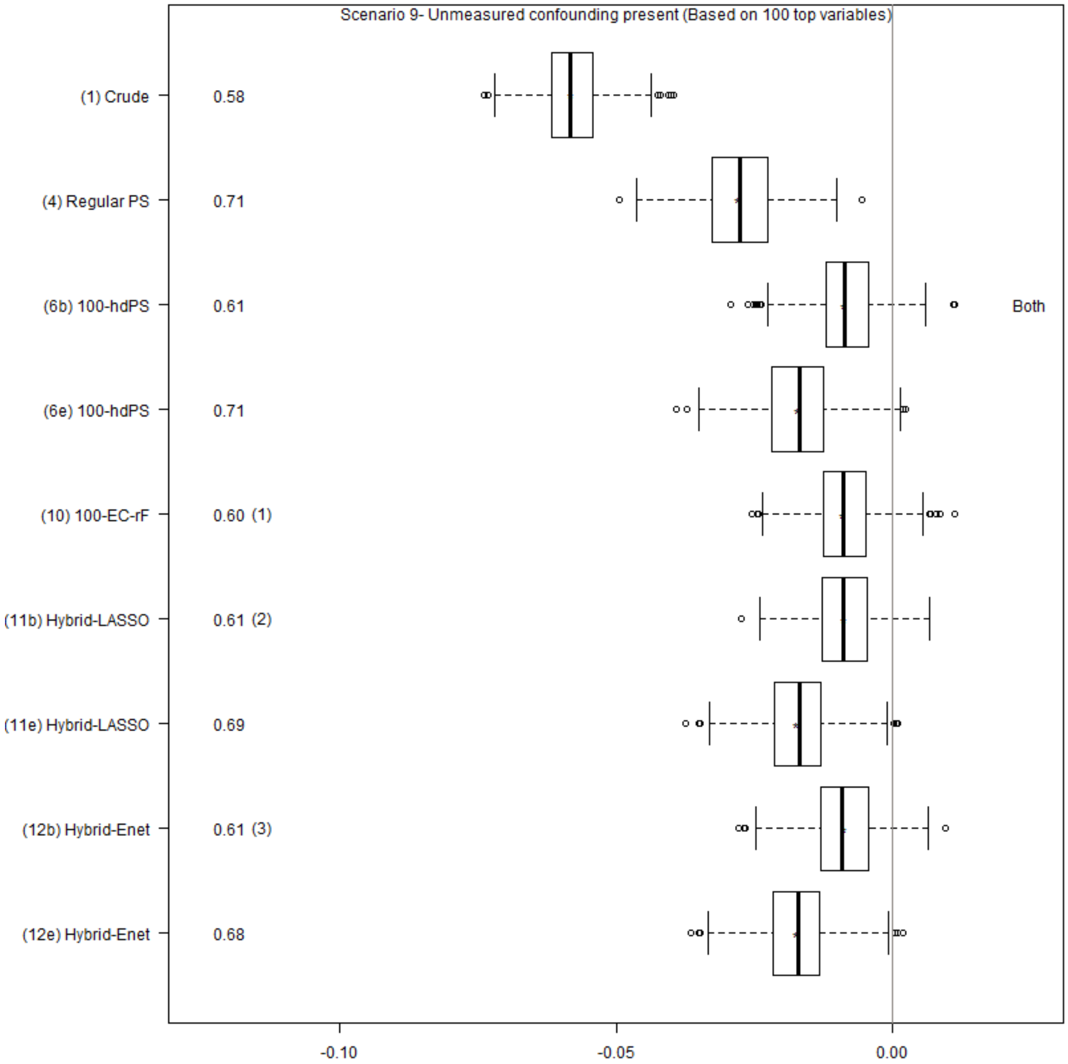
eFigure A.49: Plasmode Simulation Scenario 6-A



eFigure A.50: Plasmode Simulation Scenario 7-A

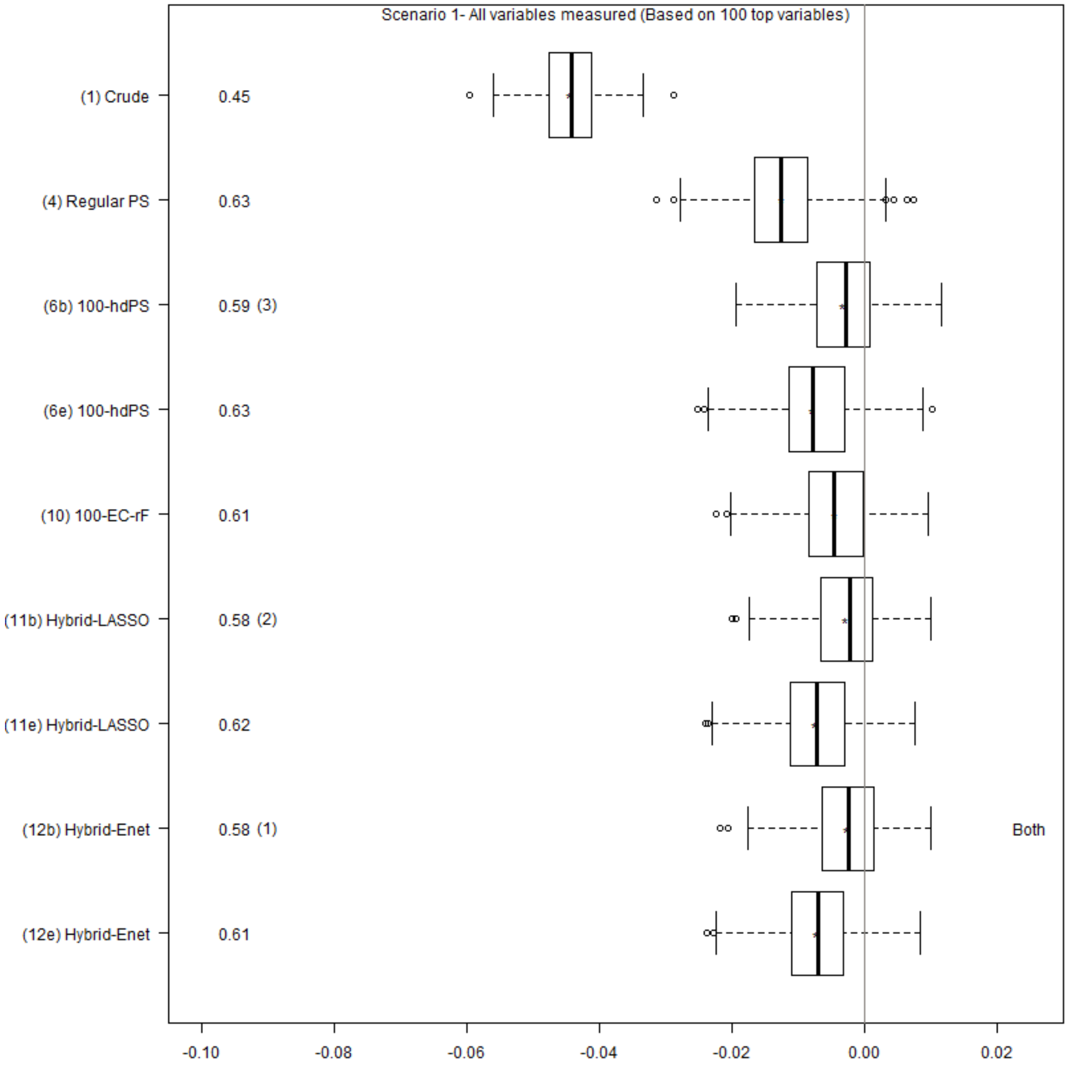


eFigure A.51: Plasmode Simulation Scenario 8-A

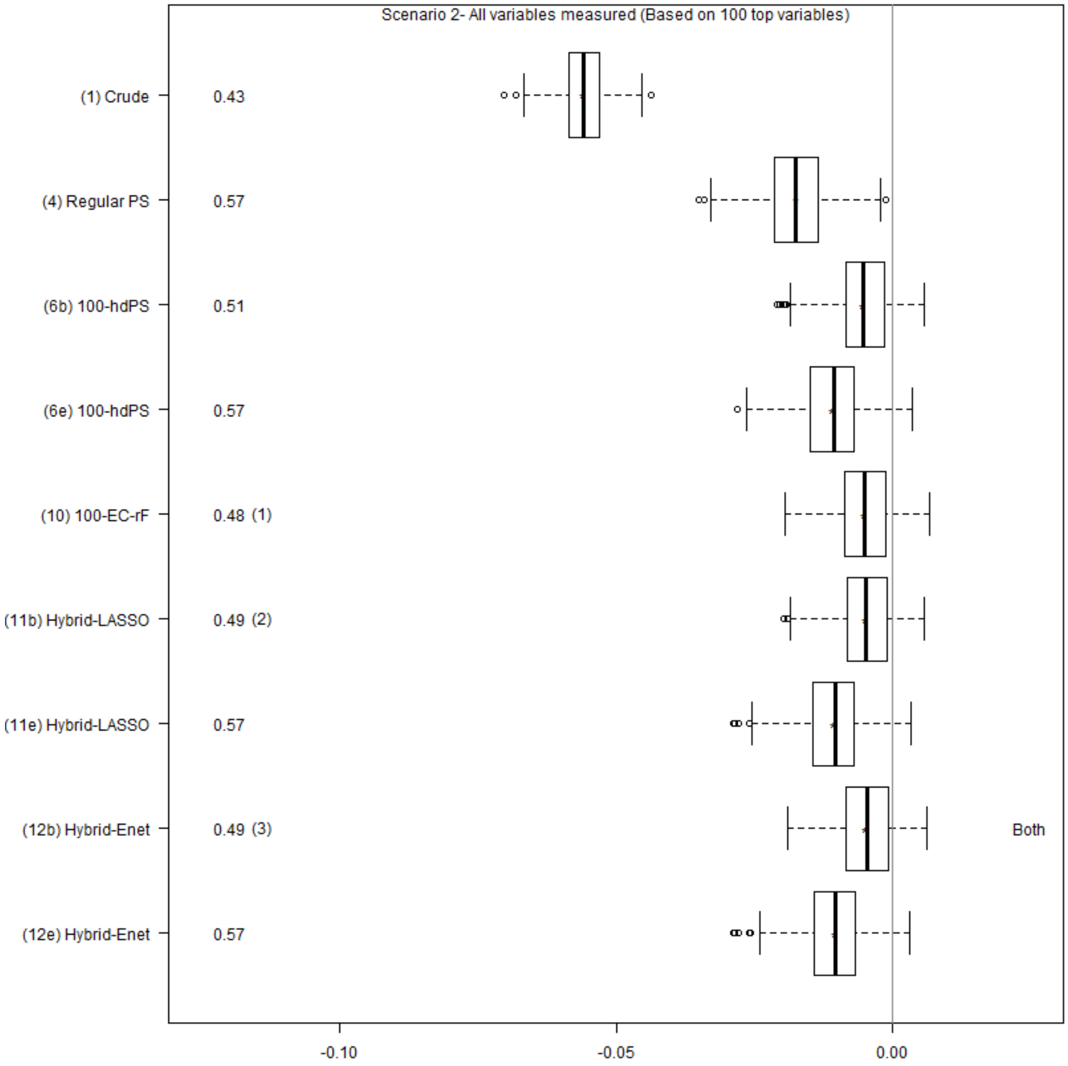


eFigure A.52: Plasmode Simulation Scenario 9-A

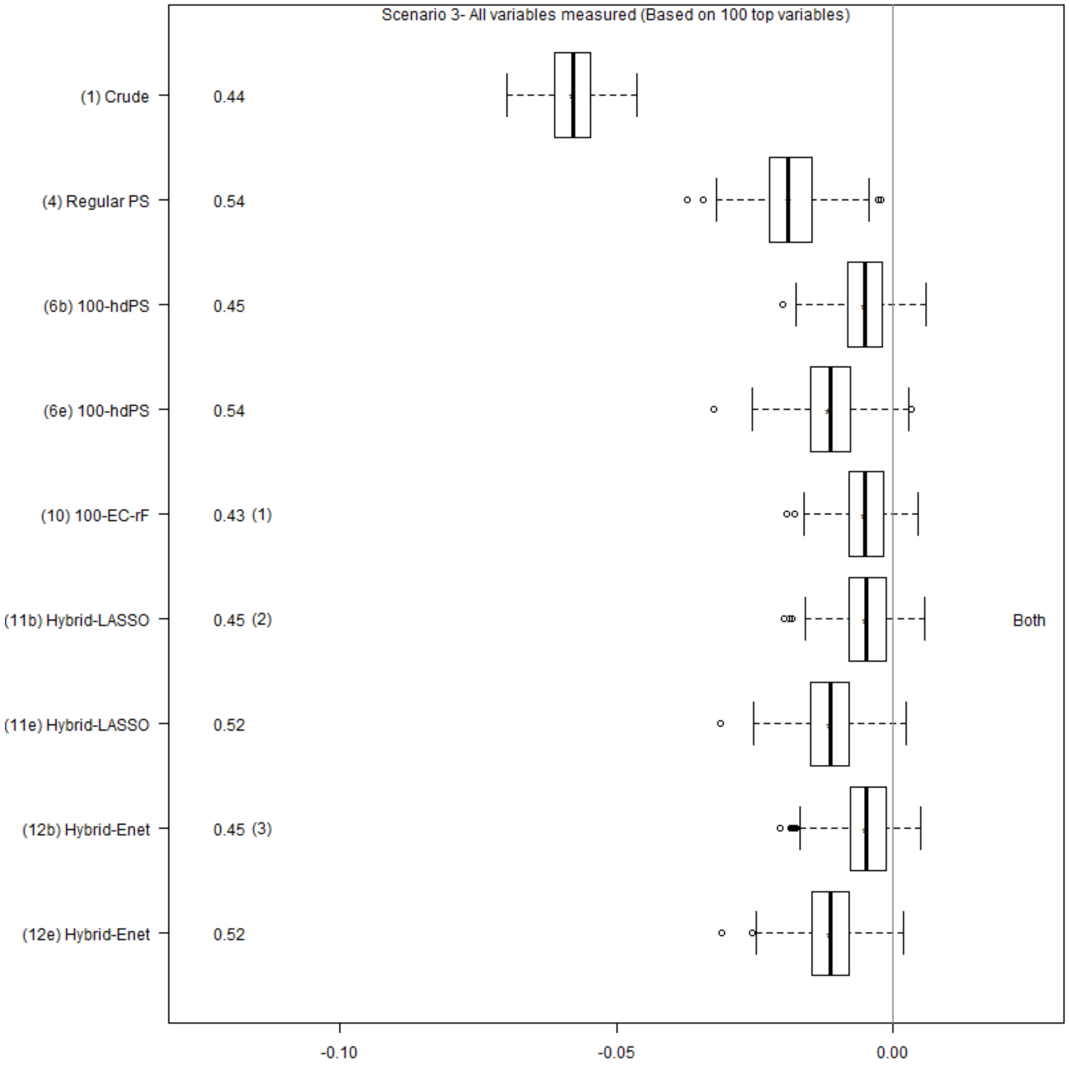
A.9.8 If all variables accounted (Based on top 100 selected variables)



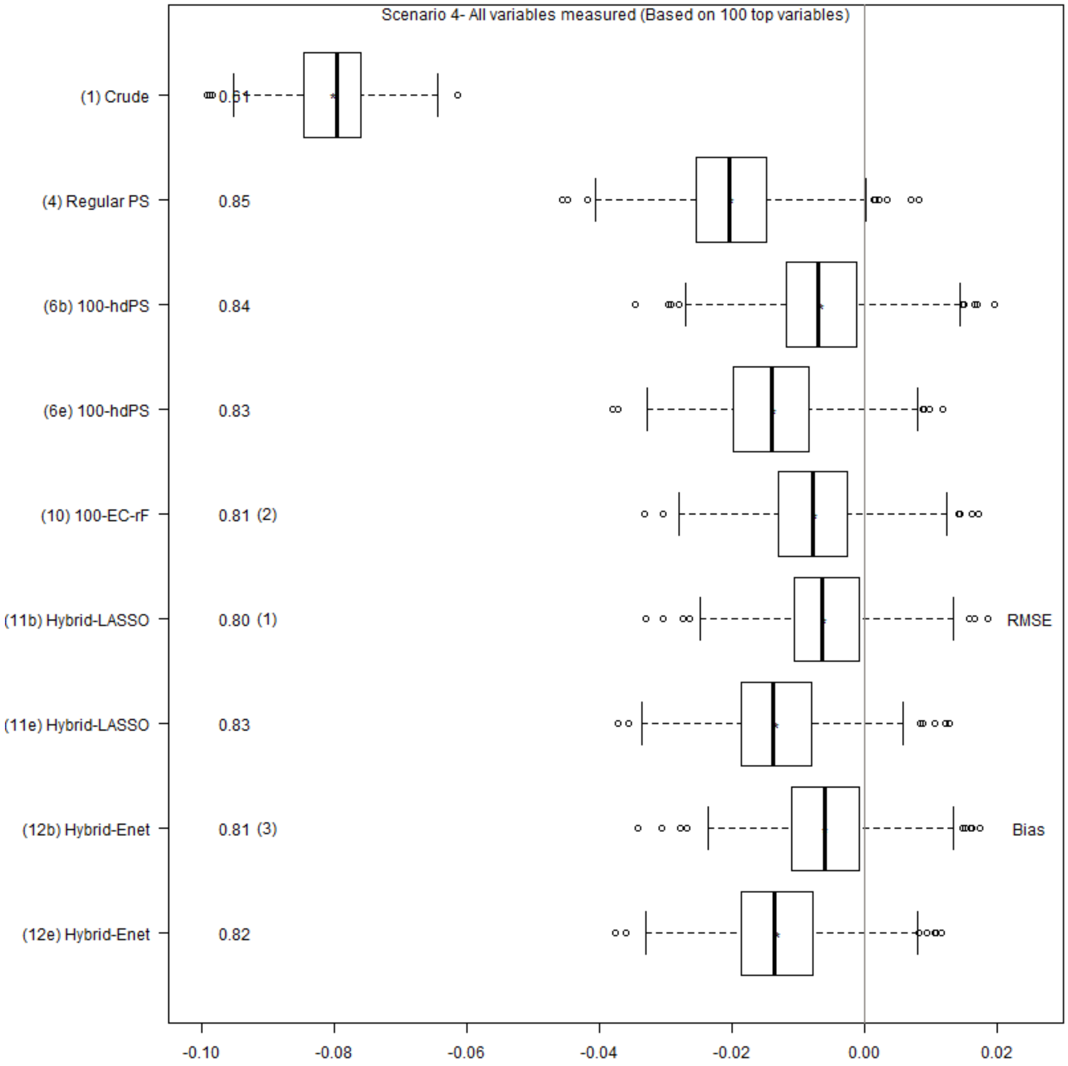
eFigure A.53: Plasmode Simulation Scenario 1-A



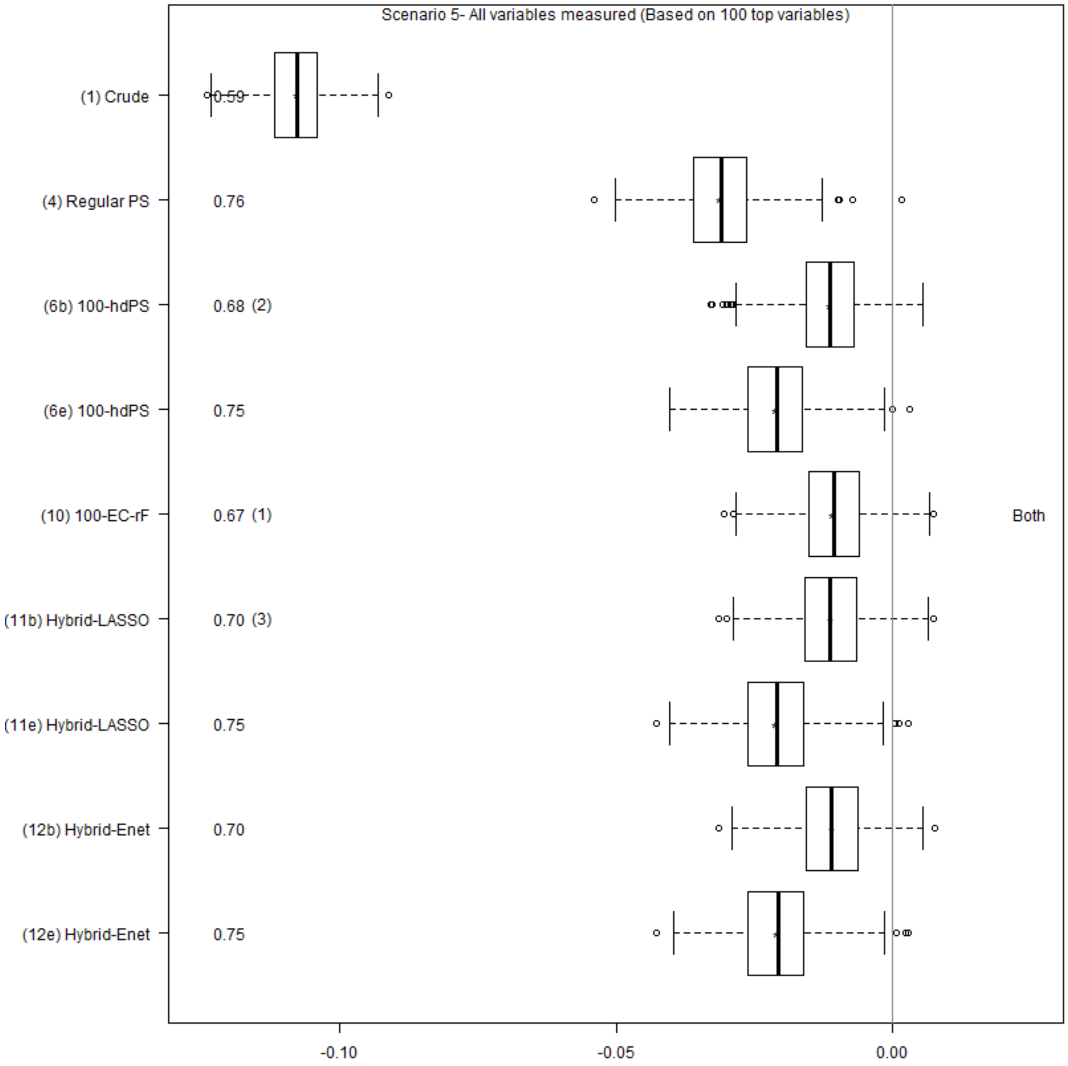
eFigure A.54: Plasmode Simulation Scenario 2-A



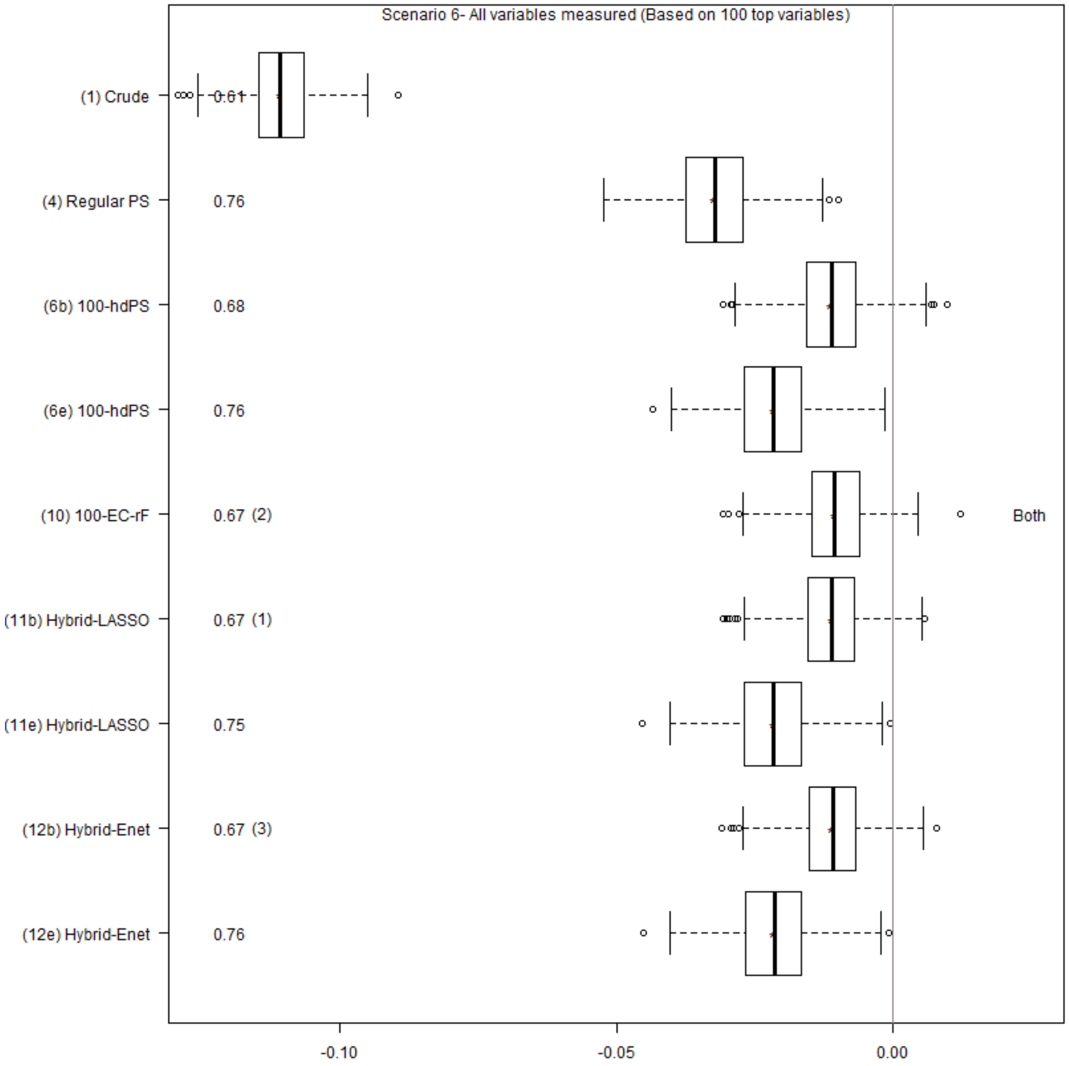
eFigure A.55: Plasmode Simulation Scenario 3-A



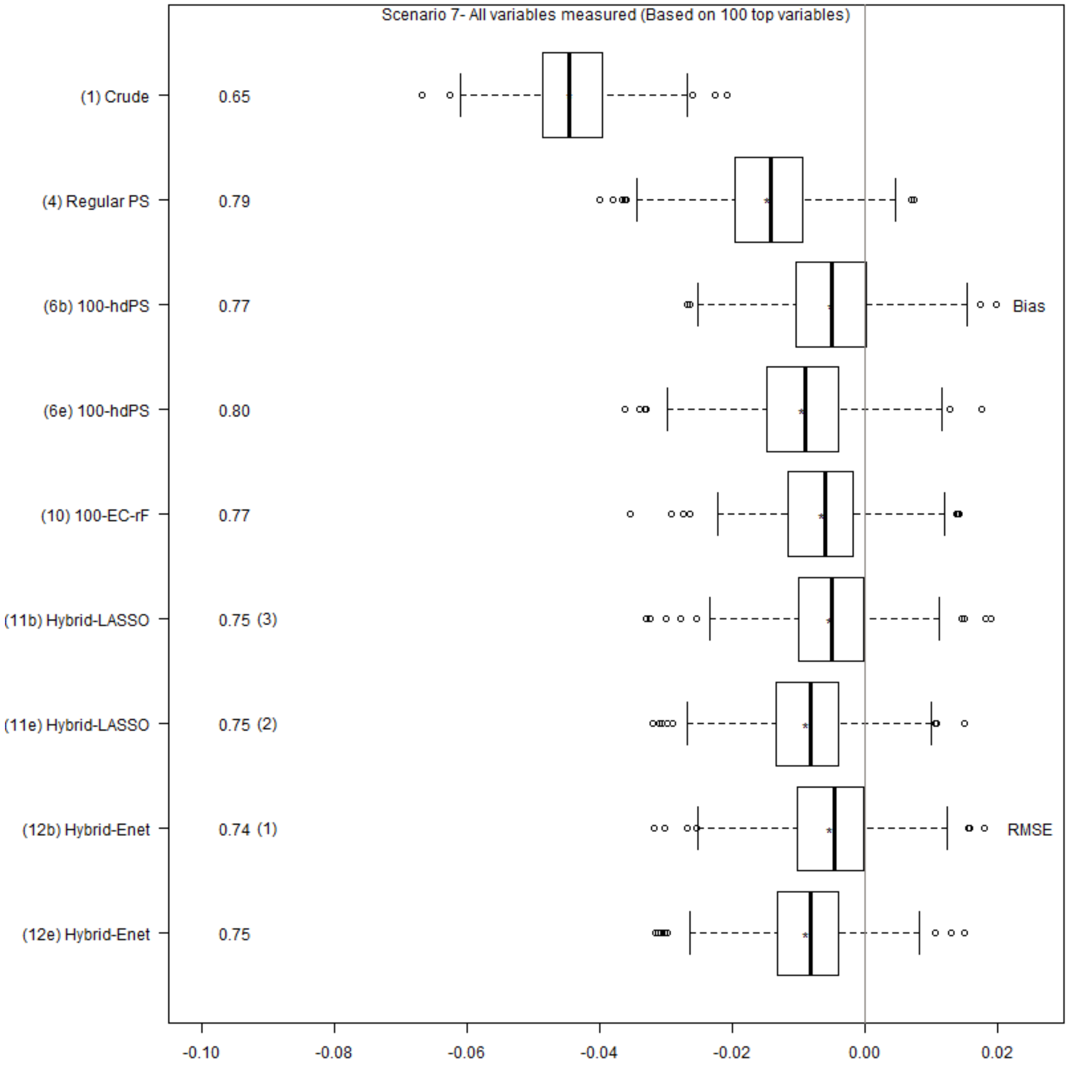
eFigure A.56: Plasmode Simulation Scenario 4-A



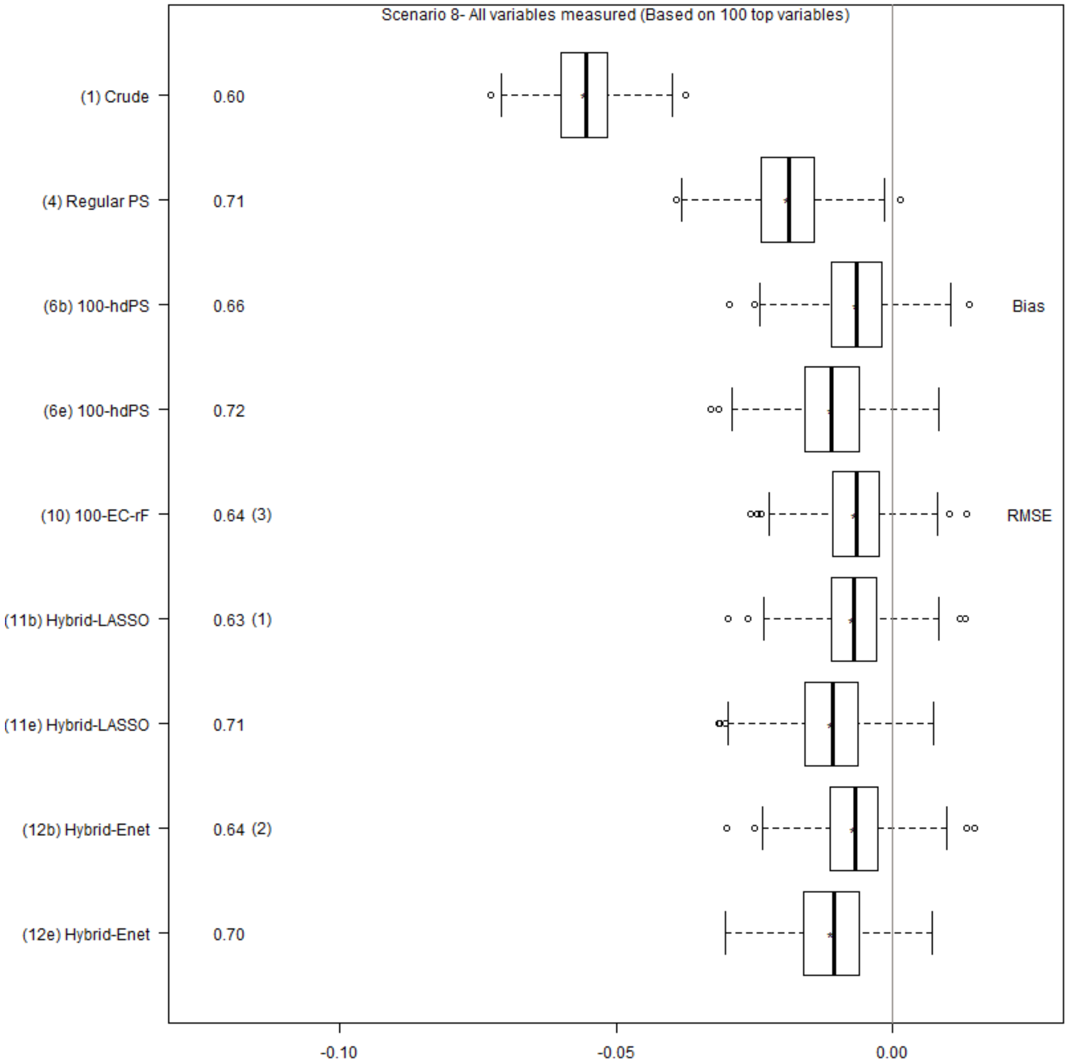
eFigure A.57: Plasmode Simulation Scenario 5-A



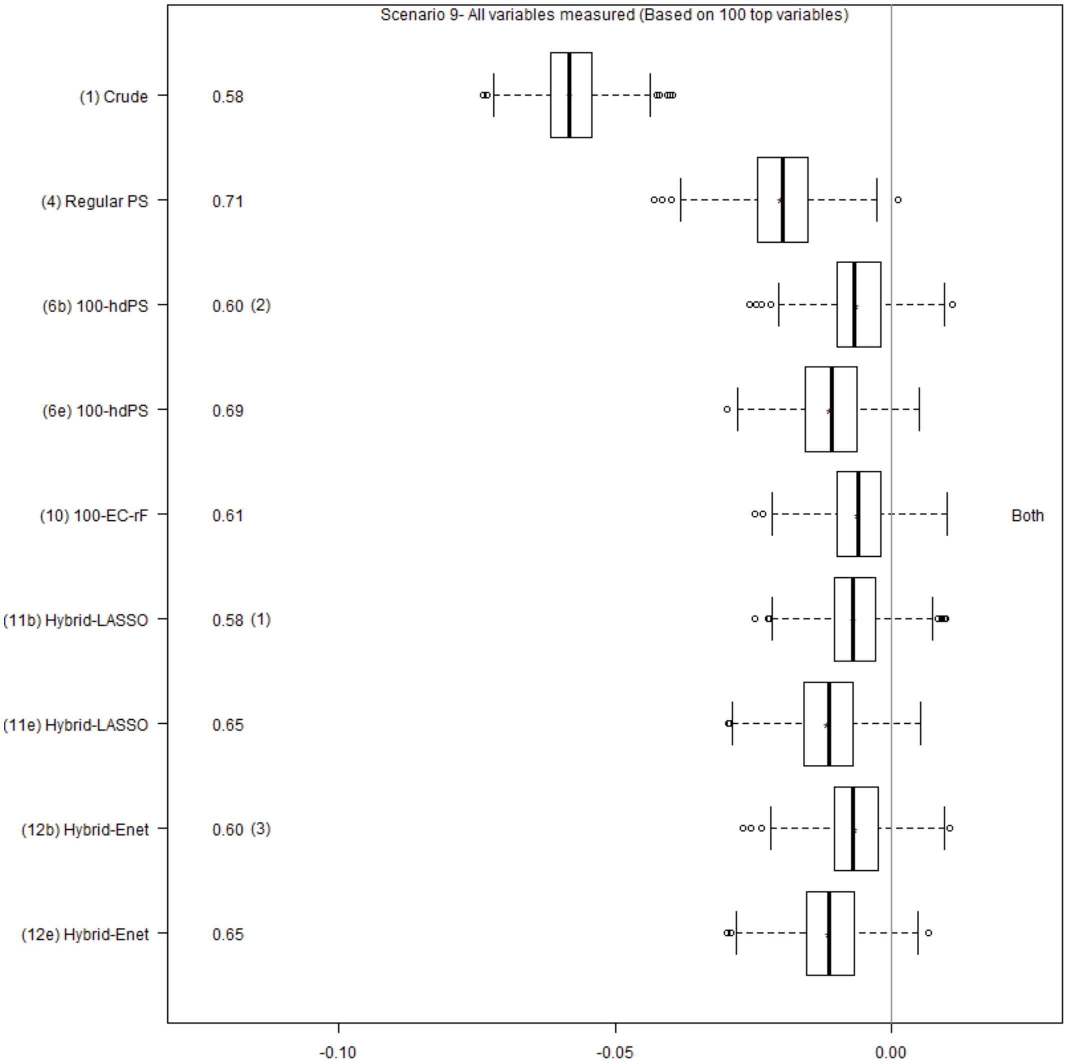
eFigure A.58: Plasmode Simulation Scenario 6-A



eFigure A.59: Plasmode Simulation Scenario 7-A

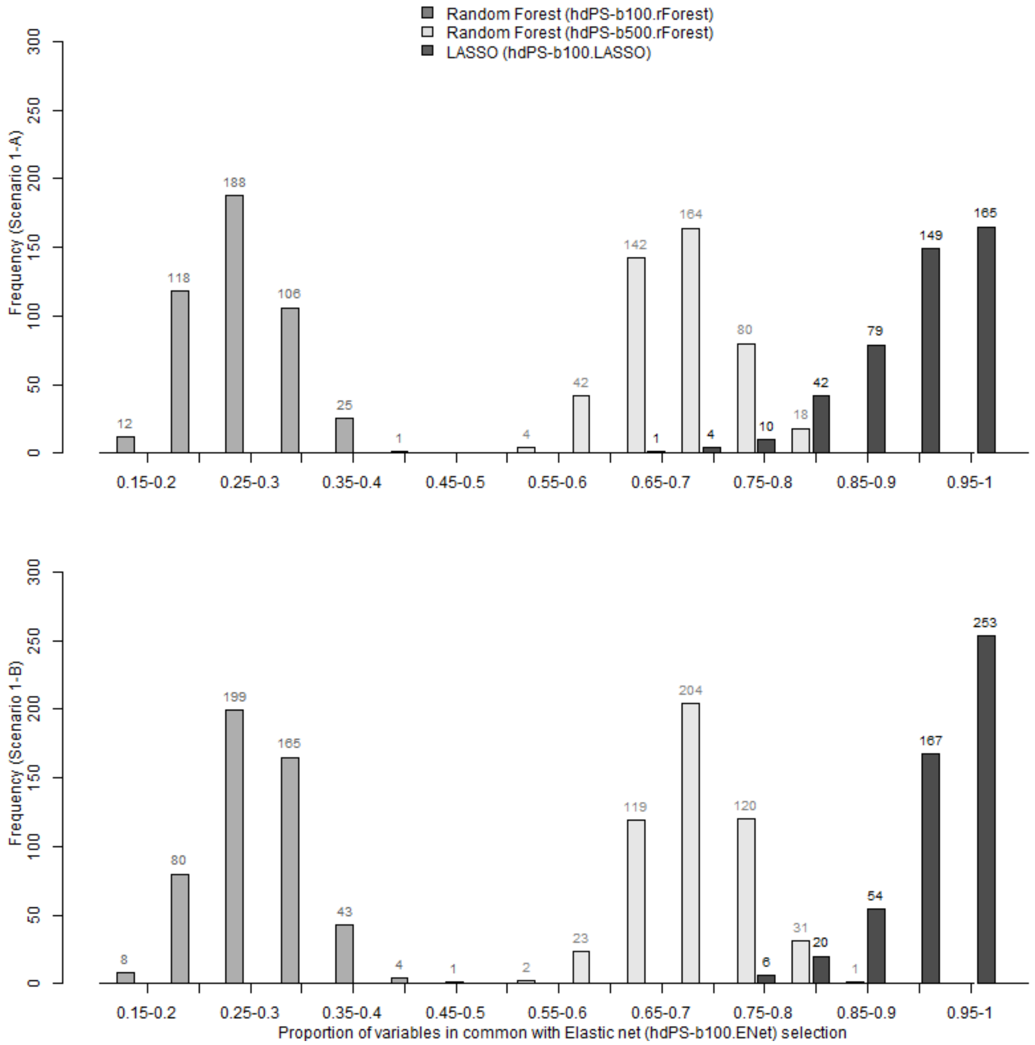


eFigure A.60: Plasmode Simulation Scenario 8-A

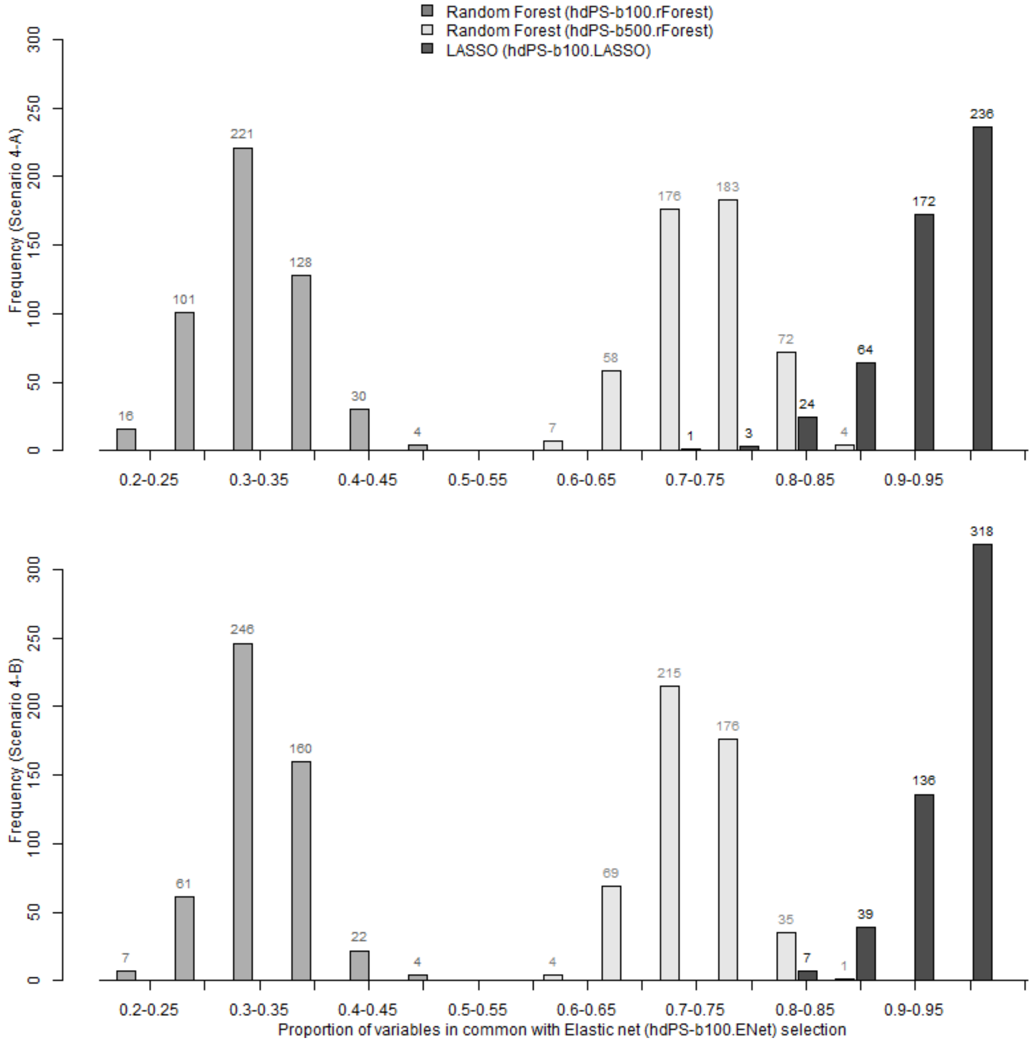


eFigure A.61: Plasmode Simulation Scenario 9-A

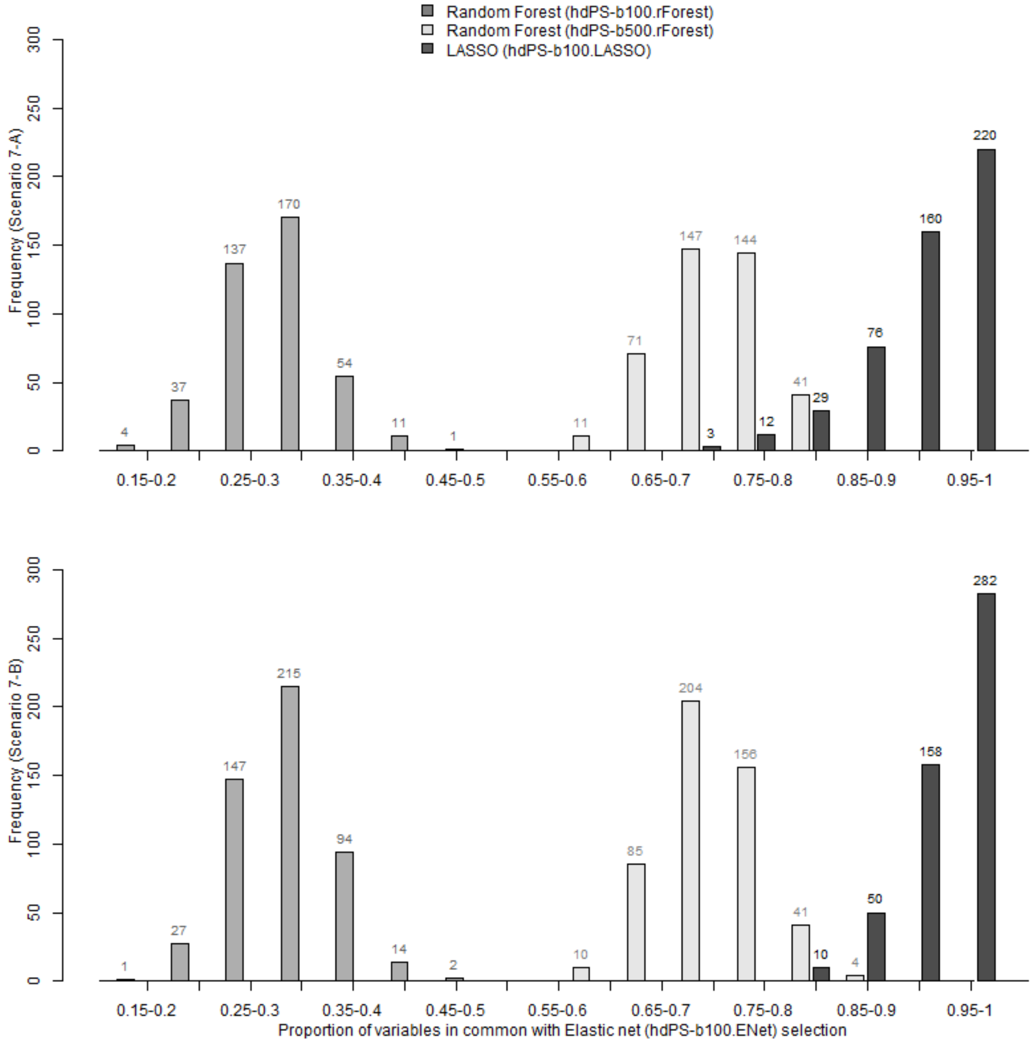
A.10 Proportion of common variables



eFigure A.62: Histogram for scenario 1



eFigure A.63: Histogram for scenario 4



eFigure A.64: Histogram for scenario 7

A.11 General Limitations of hdPS approach

Depending on how well the baseline is defined, the Bross formula may detect colliders as confounders, and adjusting for these variables may amplify bias (popularly known as "M-bias"[10]). Also, when it is difficult to determine whether a covariate is a confounder or an instrument, simulation studies in low-dimensional setting suggest that net bias will reduce if we decide to adjust for it ("Z-bias"[11]). Both of these limitations will still apply when machine learning methods are used. It is, however, argued that, in a high-dimensional setting, net bias resulting from the theoretical presence of M and Z-bias should be minimal[12].

In general, empirical-covariates are not collected for research purpose, and the interpretation is unclear[12]. Fortunately for PS-type models, the prediction is of main interest. There are many ways to utilize propensity score in the analysis, such as matching, stratification and weighting[13]. In this paper, we considered deciles of propensity scores as a covariate in the corresponding outcome analysis (as in previous literature[1], even tough this may not be the most optimal proposnsity score adjustment approach[14]). Here, propensity scores are used as a tool for data reduction. Such propensity score-type analysis is more appropriate than the regression adjustment in the high-dimensional setting we are considering here and the results from both analysis should be different, unlike the low-dimensional setting[15–17].

The hdPS analysis is a robust approach primarily to deal with residual confounding[10]. However, conceptually, this is not a straightforward extension to PS analysis. The original proposal of variable selection for the PS model was based on achieving better covariate balance. Researchers have repeatedly cautioned against the use of outcome information form the data while estimating the PSs[18–22]. However, when considering bias-based hdPS methods, we do exactly that; we rank and select empirical-covariates based on the relationship with the outcome. This criticism is also valid for machine-learning and hybrid methods; we also use information from an outcome analysis to identify important risk factors to be used later in building a PS model. Use of such information in the PS model generally prevents us from separating the design and analysis stages of a study[23,24]. However, this original proposal of relying on balance measures did assume that there all confounders are known and measured, which is a steep departure from the scenarios where hdPS analyses are generally attempted[12]. However, exposure-based hdPS are free from this criticism. Then again, exposure-based ranking scores utilize information about exposure prevalence to rank variables, and their performances are generally inferior in most settings[5].

References

[1] S. Schneeweiss, J.A. Rassen, R.J. Glynn, J. Avorn, H. Mogun, and M.A. Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512, 2009.

[2] T. Schuster, M. Pang, and R.W. Platt. On the role of marginal confounder prevalence–implications for the high-dimensional propensity score algorithm. *Pharmacoepidemiology and drug safety*, 24(9):1004–1007, 2015.

[3] Sebastian Schneeweiss, Wesley Eddings, Robert J Glynn, Elisabetta Patorno, Jeremy Rassen, and Jessica M Franklin. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology*, 2017.

[4] I.D.J. Bross. Spurious effects from an extraneous variable. *Journal of chronic diseases*, 19(6):637–647, 1966.

[5] J.M. Franklin, W. Eddings, R.J. Glynn, and S. Schneeweiss. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *American journal of epidemiology*, 182(7):651–659, 2015.

[6] J.M Franklin, Y Abdia, and S.V. Wang. 'plasmode' simulation. r package version 0.1.0, 2017. URL https://cran.r-project.org/web/packages/Plasmode/index.html.

[7] J.M. Franklin, S. Schneeweiss, J.M. Polinski, and J.A. Rassen. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis*, 72:219–226, 2014.

[8] Jessica M Franklin, Wesley Eddings, Peter C Austin, Elizabeth A Stuart, and Sebastian Schneeweiss. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Statistics in Medicine*, 2017.

[9] M. Pang, T. Schuster, K.B. Filion, M. Eberg, and R.W. Platt. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*, 27(4):570–577, 2016.

[10] M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114, 2010.

[11] J.A. Myers, J.A. Rassen, J.J. Gagne, K.F. Huybrechts, S. Schneeweiss, K.J. Rothman, M.M. Joffe, and R.J. Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222, 2011.

[12] J.A. Rassen and S. Schneeweiss. Using high-dimensional propensity scores to automate con-

founding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and drug safety*, 21(S1):41–49, 2012.

[13] Jessica A Myers and Thomas A Louis. Comparing treatments via the propensity score: stratification or modeling? *Health Services and Outcomes Research Methodology*, 12(1):29–43, 2012.

[14] Melissa M Garrido. Covariate adjustment and propensity score. *Jama*, 315(14):1521–1522, 2016.

[15] Baiju R Shah, Andreas Laupacis, Janet E Hux, and Peter C Austin. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of clinical epidemiology*, 58(6):550–559, 2005.

[16] Til Stürmer, Manisha Joshi, Robert J Glynn, Jerry Avorn, Kenneth J Rothman, and Sebastian Schneeweiss. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology*, 59(5):437–e1, 2006.

[17] Wolfgang C Winkelmayer and Tobias Kurth. Propensity scores: help or hype? *Nephrology Dialysis Transplantation*, 19(7):1671–1673, 2004.

[18] D.B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8 Part 2):757–763, 1997.

[19] Paul R Rosenbaum. Observational studies. In *Observational Studies*, pages 1–17. Springer, 2002.

[20] Donald B Rubin. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety*, 13(12):855–857, 2004.

[21] Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.

[22] E.A. Stuart, B.K. Lee, and F.P. Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S84–S90, 2013.

[23] J Michael Oakes and Timothy R Church. Invited commentary: advancing propensity score methods in epidemiology. *American journal of epidemiology*, 165(10):1119–1121, 2007.

[24] Layla Parast, Daniel F McCaffrey, Lane F Burgette, Fernando Hoces de la Guardia, Daniela Golinelli, Jeremy NV Miles, and Beth Ann Griffin. Optimizing variance-bias trade-off in the twang package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*, pages 1–23, 2016.