Manuscript

Type of manuscript: Original Article

Complete manuscript title: **Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm?**

Mohammad Ehsanul Karim[*1,2], **Menglan Pang**[3,4], and **Robert W. Platt**[3,5,6]

[1]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada

[2]Center for Health Evaluation and Outcome Sciences (CHÉOS), Providence Health Care, Vancouver, British Columbia, Canada

[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada.

[4]Centre For Clinical Epidemiology, Lady Davis Research Institute, Jewish General Hospital, Montreal, Quebec, Canada.

[5]Department of Pediatrics, McGill University, Montreal, Quebec, Canada

[6]The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada.

Details:

Abbreviated running title: Machine Learning alternatives to hdPS

Word counts for abstract: 162

Word counts for main text (excluding references): 4,080

Total number of pages: 16

Number of text pages: 13

Number of table pages: 3

Number of figure pages: 0

[*]Mohammad Ehsanul Karim, Assistant professor, School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC V6T 1Z3; and Scientist / Biostatistician, Centre for Health Evaluation and Outcome Sciences (CHÉOS), St. Paul's Hospital, 588-1081 Burrard St, Vancouver, BC V6Z1Y6, Canada. Email: ehsan.karim@ubc.ca, Tel.: +1604-682-2344 ext. 64251, Fax: +1604-806-8005, ORCiD: 0000-0002-0346-2871

Funding Information

This work was supported by a post-doctoral fellowship from the Canadian Network for Observational Drug Effect Studies (CNODES). CNODES, a collaborating centre of the Drug Safety and Effectiveness Network (DSEN), is funded by the Canadian Institutes of Health Research (CIHR). M.E.K. is a Scientist and Biostatistician at the Centre for Health Evaluation and Outcome Sciences (CHÉOS), faculty of Medicine, UBC. M.P. holds a studentship from the Fonds de Recherche du Québec - Santé (FQR-S). R.W.P. holds the Albert Boehringer I Chair in Pharmacoepidemiology, and is a member of the Research Institute of the McGill University Health Centre, which is supported by core funds from FQR-S.

Conflict of Interest

M.E.K. has received accommodation costs from the endMS Research and Training Network (2011, 2012), Statistical Society of Canada (2016) to present at conferences, and from Pacific Institute for the Mathematical Sciences (2013), the Canadian Statistical Sciences Institute (2016) to attend workshops. R. W. P. has received fees for service for consulting from Abbvie, Amgen, Eli Lilly, and Searchlight Pharma, for teaching from Novartis, and for scientific steering committee membership from Pfizer.

Availability of Data and Code for Replication

Software code hints are provided in the supporting material (as an eAppendix) for implementing the methods. Retrospective population-based cohort Dataset from the Clinical Practice Research Datalink (CPRD) is not publicly available due to patient confidentiality reasons.

Abstract

The use of retrospective healthcare claims datasets is frequently criticized for the lack of complete information on potential confounders. Utilizing patient's health status-related information from claims datasets as surrogates or proxies for mismeasured and unobserved confounders, the high-dimensional propensity score algorithm enables us to reduce bias. Using a previously published cohort study of post-myocardial infarction statin use ($1998 - 2012$), we compare the performance of the algorithm with a number of popular machine learning approaches for confounder selection in high-dimensional covariate spaces: random forest, least absolute shrinkage and selection operator, and elastic net. Our results suggest that, when the data analysis is done with epidemiologic principles in mind, machine learning methods perform as well as the high-dimensional propensity score algorithm. Using a plasmode framework that mimicked the empirical data, we also showed that a hybrid of machine learning and high-dimensional propensity score algorithms generally perform slightly better than the both in terms of mean squared error, when a bias-based analysis is used.

Keywords: *confounding; epidemiologic methods; high-dimensional propensity score; machine learning; observational data analysis.*

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.

Introduction

Observational studies are the most pragmatic means of addressing drug efficacy questions under 'real-life' clinical practice settings[1]. However, when we collect data from observational sources, the balance of covariates at baseline may no longer hold. Such imbalance could be mitigated easily by adjusting for respective confounders in a regression model or in a propensity score[2,3] context. However, these methods assume "no unmeasured confounding"[4,5] i.e., a sufficient set of confounders are recorded and adjusted in the analysis, either directly or through the propensity score.

Observational studies of drug efficacy often use administrative datasets. These datasets are not primarily collected for research purposes, so the investigators do not have much control over what covariates are measured. Therefore, studies based on pharmacoepidemiologic healthcare claims databases are frequently criticized for the lack of complete information on

the potential confounders[6]. Historically, researchers have adjusted for the set of available and measured covariates via regression or propensity score adjustments. When researchers perform data analysis and adjustment using only measured confounders, the estimated treatment effects may be biased and subject to residual confounding[7].

Fortunately, a wide range of health care utilization databases routinely collect a large volume of digital electronic administrative records. These data sources additionally contain longitudinal information about patients' health status and various related information, such as unique medical diagnoses, procedures, providers, health insurance plans, and prescription dispensing, as well as information from electronic medical records, laboratory results, accident registries, etc. This information, usually in the form of codes that can be translated into thousands of variables, are potentially correlated with the important unmeasured or imprecisely measured confounders[5,8] and thus, can be used as overall proxies of them[9].

As these data are not usually collected for research purposes, it is not clear how to make optimal use of such information in an analytic setting. Conventional pharmacoepidemiological studies do use diagnosis, procedure and drug prescription to define their exposure, outcome and covariates of interests, but they do not consider all the information. Schneeweiss and his colleagues[10] introduced an algorithm called the high-dimensional propensity score algorithm that advocated the use of all the information available in health care claim data. Since its publication, there has been a growing interest in this approach (see eFigure A.1).

Unlike typical pharmacoepidemiologic studies, considering such a massive amount of proxy data is essentially a big data problem[11]. According to the epidemiologic literature, in our propensity score model, we need to include variables associated with the outcome, even if they are seemingly unrelated to the treatment decision[12]. Using 'kitchen-sink' models that indiscriminately adjust for all proxy covariates without considering how they affect treatment and outcome may be counterproductive in terms of reducing bias or obtaining an efficient estimate of the treatment effect[2,12]. This is particularly the case for instrumental variables because adjusting for such variables may amplify bias and increase variance[13]. However, variable selection for confounder adjustment in this high-dimensional setting is a challenging problem because hand-picking such covariates (e.g., by an expert) is not practical. The proposed high-dimensional propensity score algorithm offers a practical way to select a large

number of covariates that are suitable for the propensity score model. This algorithm automates the selection of adjustment covariates in seven well-defined steps[10] by empirically assessing bivariate associations between the proxy variables and outcome variables, adjusting for exposure prevalence. Based on their potential for confounding (usually measured via a bias score[14]), variables are assigned a rank (prioritized), and only the highest ranked variables are selected for inclusion in a propensity score analysis (bias-based ranking). Generally, the 100 or 500 top-ranked empirical covariates are selected. These ranked empirical covariates are known as "high-dimensional propensity score variables"[6]. In simulation and empirical studies[10,15,16], the high-dimensional propensity score algorithm has been shown to optimally reduce bias in many comparative effectiveness studies. Other criteria such as exposure based ranking, are also suggested in the literature for situations with few exposed outcomes[15] (eAppendix A.5 for corresponding formulas).

To deal with the challenge of dimensionality, many machine-learning methods have been proposed in the statistical, epidemiologic as well as big-data literature[17,18]. Methods based on, say, classification and regression trees[18] are inherently flexible, data-adaptive and associated with less strict assumptions, and have considerable potential to capture various features of the data, such as nonlinear patterns, interaction, and higher-order effects[19–23]. Most of these machine-learning methods, however, tend to focus on increasing the predictive accuracy[24]. Similar to the previously discussed propensity score settings, blindly including all possible covariates into the machine-learning methods may amplify bias in the analysis due to the inclusion of covariates that are irrelevant to the outcome[13,25].

In this work, we deal with customizing some of these machine-learning methods to incorporate the appropriate variables that follow the epidemiologic principles (e.g., include variables associated with the outcome in the propensity score modeling). As the title suggests, the current research aims at finding out whether the machine-learning Methods, trained with the relevant epidemiologic principles[12], can outperform the high-dimensional propensity score algorithm. We will also consider hybrid approaches to bring together both types of algorithms and harness their respective strengths.

Methods

Empirical Dataset

We utilized a retrospective population-based cohort study using the United Kingdom data from the Clinical Practice Research Datalink (CPRD)[26]. These data were linked to the Hospital Episode Statistics database (HES), which contains detailed hospitalization records. A total of 32,792 patients aged 18 and older, and diagnosed with an initial post-myocardial infarction (MI) were drawn from the databases between 1 April 1998 and 31 March 2012. This cohort consists of 19,121 patients treated with statin within 30 days after the diagnosis of MI. All-cause mortality was evaluated as any death recorded in the databases during the one-year follow-up period. Previous research identified five important confounders: age, sex, obesity, smoking, and history of diabetes[26]. Twenty-four other potential known confounders were designated as predefined covariates for the study (listed in the eAppendices A.2 and A.3). From four linked data dimensions, we create binary proxy covariates, following the high-dimensional propensity score algorithm[10], considering the top 200 most prevalent codes (details in eAppendix A.4). To distinguish these covariates from the investigator specific covariates, we call them empirical covariates[27]. This study was approved by the Independent Scientific Advisory Committee for Medicines and Healthcare Products Regulatory Agency database research (protocol number 14_018) and the Research Ethics Board, Jewish General Hospital, Montreal, Canada.

Adjustment Tools

We have listed the high-dimensional propensity score methods and the machine-learning alternatives under consideration in Table 1. In this work, we used deciles of the propensity score distribution as a covariate in the outcome analysis. We calculate the odds ratio (OR) from methods (1-5) for comparison purposes. Approaches (6-7) are high-dimensional propensity score methods. Pure machine-learning methods, such as least absolute shrinkage and selection operator (LASSO) (8), were recently proposed as an alternative to the high-dimensional propensity score algorithm[6] (shown via data analysis and simulation). We propose to use a machine-learning approach in this work known as elastic net (9)[28]. This approach is capable to generally selecting a more stable superset of the LASSO selected confounders[29]. Random forest method (10)[30] is another machine-learning approach that has

been recently used in another data analysis (but not simulation) context in comparison to high-dimensional propensity score[31]. This approach uses a prediction error based criterion to decide the 'variable importance' of each empirical covariate in predicting the outcome. We also consider hybrid approaches (11-12), that combine high-dimensional propensity score and machine-learning approaches[6,31]. eAppendix A.6 discusses the technical details of the software use.

Plasmode simulation

To evaluate the performance of the high-dimensional propensity score algorithm and the machine-learning methods in a realistic high-dimensional covariate settings, we conducted a plasmode simulation[16,32] study mimicking our empirical study where associations and correlations between covariates reflect real-world settings (details in eAppendix A.7).

Simulation Specifications

In total, we considered 18 plasmode simulation settings (parameter specifications are listed in Table 2). These settings fall under two broad scenarios: (U-set) unmeasured confounding present, i.e., all variables (empirical as well as investigator-specified) were used to generate data, but five important confounders were omitted during data analysis, which is the general scenario where analysts are more likely to engage a high-dimensional propensity score analysis, and (A-set) all variables are measured and included in the analysis. Simulation settings are varied by the true underlying model generating the outcome, assigned covariate effect multiplier (γ), the prevalence of outcome and exposure (p_Y and p_E respectively), and the presence of unmeasured confounding, all of which has been identified as useful parameters for plasmode simulations[16,33,34]. For simplicity, we set the true odds ratio to be 1. To avoid the problem of noncollapsibility of the odds ratio[35,36], we followed the usual practice in the literature to estimate a measure of effect that is collapsible (e.g., risk difference)[32,37] and calculate bias and mean squared error (MSE) accordingly (considering risk difference of zero to be true parameter). In each of these simulation scenarios, we considered generating $N = 500$ datasets with $m = 10{,}000$ subjects in each dataset.

Table 1: Adjustment Tools under consideration: (a) Basic comparators, (b) High-dimensional propensity score methods, (c) Machine learning methods and (d) Hybrid approaches.

Name	Description
Basic comparators	
(1) Crude	Crude analysis without any covariate adjustment
(2) PS Important	PS analysis with only the 5 important covariates.
(3) Regression	Regression adjustment with 29 investigator-specified covariates.
(4) Regular PS	PS analysis with 29 investigator-specified covariates.
(5) kitchen-sink	All empirical covariates (ECs) as well as 29 investigator-specified covariates are placed in the PS model without any hdPS or machine-learning pre-selection.
hdPS	
(6) 500-hdPS	PS analysis with 500 hdPS variables.
(7) 100-hdPS	PS analysis with 100 hdPS variables.
Pure machine-learning[a]	
(8) All-EC-LASSO	(a) Initialize the outcome model with all possible ECs. In this model, based on the relationship with the outcome, LASSO shrinks a number of unstable estimated covariate coefficients to zero and eliminates the respective hdPS covariates from the outcome model. LASSO will return a reduced model with a subset of ECs that are meaningfully associated with the outcome. (b) We then use this subset of ECs to build our PS model and (c) subsequently perform outcome analysis again using the treatment and PS deciles as covariates.
(9) All-EC-Enet	Similar to approach (8), but using elastic net instead of LASSO.
(10) 500-EC-rF	(a) Based on the outcome and covariate association in a multivariate random forest model, 500 top important variables are identified and (b) they are used to build a reduced PS model. (c) Subsequent outcome-exposure association are then assessed after adjusting for the estimated PS deciles.
Hybrid[b]	
(11) Hybrid-LASSO	LASSO models will perform variable selection on the selected 500 hdPS variables and reduce the number of covariates to be used in the PS model.
(12) Hybrid-Enet	Similar to approach (11), but using elastic net instead of LASSO.

PS, propensity score; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; EC, empirical covariate.

[a] All ECs are entered into in the machine-learning algorithms. No high-dimensional propensity score pre-selections are necessary.

[b] Only the 500 ECs selected by the high-dimensional propensity score algorithm are entered into the initial model.

Table 2: Plasmode simulation settings under consideration modifying the cohort of post-myocardial infarction statin use ($1998-2012$)

Simulation Scenario[a]	γ[b]	p_E[c]	p_Y[d]	Unmeasured confounders
1-U	1	40	5	Yes[e]
2-U	3	40	5	Yes
3-U	5	40	5	Yes
4-U	1	40	10	Yes
5-U	3	40	10	Yes
6-U	5	40	10	Yes
7-U	1	10	5	Yes
8-U	3	10	5	Yes
9-U	5	10	5	Yes
1-A	1	40	5	No
2-A	3	40	5	No
3-A	5	40	5	No
4-A	1	40	10	No
5-A	3	40	10	No
6-A	5	40	10	No
7-A	1	10	5	No
8-A	3	10	5	No
9-A	5	10	5	No

[a] Each of these scenarios was generated from the following plasmode simulation's outcome generating equation: $logit\big[Pr(Y = 1)\big] = \alpha_0 + \theta \times \alpha_1 T + \gamma \times \alpha_2 X$. Here, Y is the outcome, T is the treatment indicator, X is the matrix of investigator-specified and empirical covariates. α_0 is the intercept, α_1 and α_2 are the treatment effect and the covariate effects respectively, θ and γ are multipliers of the treatment effect and the covariate effects respectively. See eAppendix A.7 for details.

[b] γ, the covariate effect multiplier, uniformly amplifies observed association between each covariate and the outcome.

[c] p_E is the prevalence of exposure.

[d] p_Y is the prevalence of outcome.

[e] discarding five most important confounders identified in previous research: age, sex, obesity, smoking, history of diabetes.

Results

Empirical Data Analysis Results

Treatment effect estimation

All our OR estimates are plotted in eFigure A.6 with corresponding confidence intervals. Without any adjustment, the estimated crude OR is 0.3. The five most important covariates are not well-balanced at baseline (see eAppendix Table A.1). Adjusting for these five important covariates in the propensity score model resulted in an estimated OR of 0.47. Considering 24 additional investigator-specified covariates (see eAppendix A.2) resulted in an OR of 0.62. Adding more covariates moved the OR to some extent. As we are dealing with observational data sources, unmeasured confounding is a concern. To reduce the effect of potential residual confounding in the analysis, researchers would prefer to adjust for more variables. Under the assumption that the collected empirical covariates are likely associated with unobserved confounders and can be used as proxies of the unmeasured confounders, we built a propensity score model with all possible empirical covariates as well as 29 investigator-specified covariates ('kitchen-sink' approach). The resulting OR is .65, which is very close to the previous OR 0.62 that we obtained from utilizing only the investigator specified covariates.

For any propensity score (and high-dimensional propensity score) model building process, it is essential to assess the balance in the propensity score distribution[38]. When we had only 29 investigator specified covariates, the propensity score from both exposure group had sufficient overlap (see eFigure A.4). However, when we included all possible empirical covariates in our kitchen-sink model, control status (0) and exposure status (1) are almost perfectly predicted by the large propensity score model, and hence there is not sufficient overlap in the middle, suggesting that the kitchen-sink model does not sufficiently adhere to the diagnostic criteria (e.g., overlap) recommended for assessing balance.

However, after selecting top 100 high-dimensional propensity score variables, we can see there is sufficient overlap in the propensity score in both groups. Even when we selected the top 500 high-dimensional propensity score variables, the overlap seems satisfactory (see eFigures A.4-A.5). The OR from the 100-high-dimensional propensity score approach is 0.74

(see eFigure A.6). When we considered more variables, say 500 high-dimensional propensity score variables, the OR is 0.78.

Using the LASSO selection model, we can get a subset of empirical covariates. Using them in propensity score analysis, we get the OR 0.76. When elastic net is used for variable selection, it results in OR of 0.77. We can see that with 500 important empirical covariates chosen by the random forest approach, the OR is 0.79. Hybrid approaches also resulted in similar ORs.

Sensitivity analysis

We performed a sensitivity analysis to check whether the use of empirical covariates in the analysis can somewhat compensate for the omitted confounders. Let us assume that we have not collected five confounders that were deemed important for this study previously[26] and we want to investigate if high-dimensional propensity score analysis can compensate for such missing or omitted information. We performed high-dimensional propensity score analysis all over again without those five confounders; results are plotted in eFigure A.7. We can see that ORs estimated the 500-high-dimensional propensity score, machine-learning methods and hybrid approaches are apparently higher than that from the propensity score analysis that included those five confounders (marked by the grey line at OR = 0.62). Therefore, methods utilizing these surrogate variables that are potentially associated with the unmeasured confounders, resulted in increasing the ORs (all ORs above 0.62). However, none of the estimates reached the same level as the earlier analyses, when we included these five confounders (compared to eFigure A.6, either of the dotted lines).

Simulation Results

If unmeasured confounding present

All the simulation results shown in graphs are sorted in the same order the approaches were presented in Table 1. Figures A.8 - A.10 demonstrate the performance of each of the approaches under consideration for simulation scenarios 1-U, 4-U, and 7-U when unmeasured confounding present (set-U) and high-dimensional propensity score variables are selected based on bias score. In all of these scenarios, all of the approaches using empirical covariates (even the 100-high-dimensional propensity score approach) performs better than the regular

propensity score approach.

When we have a higher exposure prevalence ($p_E = 40$) but a less prevalent outcome ($p_Y = 5$) in simulation scenario 1-U, in general, these approaches are associated with least bias. Bias was slightly increased when exposure prevalence was lower ($p_E = 40$ and $p_E = 10$ in scenario 4-U), but most biases are related to scenarios when outcome prevalence ($p_E = 10$ and $p_E = 5$ in scenario 7-U) is more. In all of these settings, hybrid methods (Hybrid-Enet and Hybrid-LASSO) seem to do better in terms of MSE than any of the pure machine-learning or high-dimensional propensity score algorithms. Except for scenarios 5-U and 6-U, hybrid methods continue to perform well when we consider stronger covariate associations (eAppendix A.9.2: eFigures A.11 - A.16). In those two scenarios, pure machine-learning method 500-EC-rF performs best in terms of both bias and MSE.

When high-dimensional propensity score variables were selected based on exposure-based selection in the same set-U scenarios, machine-learning methods (All-EC-Enet, 500-EC-rF and All-EC-LASSO) perform better in all situations, considering MSE as a criterion for comparison; see eFigures A.17-A.25). Note, however, that estimates obtained from high-dimensional propensity score, machine-learning methods and hybrid approaches utilizing the empirical covariates were not much different in any of the settings we have considered in terms of the magnitude of difference in the effect estimate. Considering fewer variables in the analysis did not change the results in general (see eAppendix A.9.6).

If all relevant variables are measured

In an unlikely scenario, when all relevant variables are measured and included in the analysis (set-A), hybrid methods perform well in all scenarios when bias score was used for ranking (see eFigures A.26-A.34). Again, pure machine-learning methods perform well when exposure-score was used for ranking (see eFigures A.35-A.43). Table 3 lists all the best approaches based on the chosen criteria (bias or MSE).

Table 3: Methods performing best in various simulation scenarios in terms of mean squared error and bias criteria: pure high-dimensional propensity score methods are marked as italic and pure machine-learning approaches as bold.

	Bias-based		Exposure-based	
Scenario	MSE	Bias	MSE	Bias
1-U	Hybrid-Enet	Hybrid-Enet	**All-EC-Enet**	**All-EC-Enet**
2-U	Hybrid-LASSO	*500-hdPS*	**All-EC-Enet**	**All-EC-Enet**
3-U	Hybrid-LASSO	*500-hdPS*	**All-EC-Enet**	**All-EC-Enet**
4-U	Hybrid-Enet	Hybrid-Enet	**500-EC-rF**	**500-EC-rF**
5-U	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
6-U	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
7-U	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
8-U	Hybrid-Enet	**500-EC-rF**	**All-EC-LASSO**	**All-EC-LASSO**
9-U	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
1-A	Hybrid-LASSO	**All-EC-LASSO**	**All-EC-Enet**	**All-EC-LASSO**
2-A	Hybrid-LASSO	Hybrid-LASSO	**All-EC-Enet**	**All-EC-Enet**
3-A	Hybrid-Enet	Hybrid-LASSO	**All-EC-LASSO**	**All-EC-Enet**
4-A	Hybrid-LASSO	**All-EC-Enet**	**All-EC-Enet**	**All-EC-Enet**
5-A	Hybrid-LASSO	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
6-A	Hybrid-Enet	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
7-A	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
8-A	Hybrid-Enet	**500-EC-rF**	**All-EC-LASSO**	**All-EC-Enet**
9-A	Hybrid-LASSO	Hybrid-Enet	**All-EC-LASSO**	**All-EC-Enet**

Considering 'bias' as a criterion

As shown in Table 3, the superior performance of the machine-learning or hybrid approach is also true when we consider bias as a measure of criterion instead of MSE. 500-high-dimensional propensity score performed best when the bias-based analysis was conducted in four scenarios (in set-U) and only once (in set-A) in the absence of unmeasured confounding. In terms of exposure-based analysis, 500-high-dimensional propensity score or any of the hybrid approaches never came out on top in either criterion (bias or MSE). Considering exposure-based analysis, pure machine-learning methods are always the best no matter which criterion you choose.

Proportion of chosen variables in common

Previously it was shown via simulation, that the variables chosen by the LASSO approach were mostly different than the variables selected by the bias score (bias-based and exposure-based)[6]. Apparently, the empirical covariates selected by the random forest method are also very different than the empirical covariates selected by the LASSO and elastic net method. In our data analysis context, only about 30% variables are in common when we picked 100 variables from the random forest and 100 high-dimensional propensity score variables from the elastic net approach (see eAppendix A.10 for scenarios 1, 4 and 7). However, although the variables were different, the resulting ORs were in close proximity.

Discussion

Application of machine-learning methods in the analysis of high-dimensional health care databases is not new[6,31,34,37,39,40]. Unlike much of the previous literature in this context, we clearly distinguish between high-dimensional propensity score, machine-learning, and hybrid approaches and compared them under an unified framework. One of the novel aspects of the current work is that we have utilized machine-learning for identifying and selecting confounders (based on the association between the covariates and the outcome) instead of using them in direct exposure modeling to enhance prediction, as was done in some earlier works[19,31,41]. In this paper, in the context of analyzing a healthcare administrative dataset, under the same framework, we aimed to assess the performances of three machine-learning

methods as well as two hybrid approaches (combination of high-dimensional propensity score and machine-learning methods) and compared them with a regular high-dimensional propensity score analysis in adjusting for residual confounding.

We compared high-dimensional propensity score and machine-learning methods in a retrospective cohort study of statin use post-MI and the 1-year risk of all-cause mortality. When considering 500 or more empirical covariates, the estimated ORs were between 0.76 and 0.79. Such findings are consistent with the previous study results based on the same empirical dataset using a double robust estimation approach[42]: the reported OR was 0.77 and a sensitivity analysis suggested an OR of 0.8 when the estimated propensity score were truncated at the 1st and 99th percentile to avoid creating extreme inverse probability weights[26]. However, from the analysis of observational data, no matter how sophisticated the estimation approach is, we can't be sure that we have obtained the right answer. Non-collapsibility of OR further prevents us from making meaningful comparison between ORs estimated from various approaches under consideration. Therefore, we have conducted plasmode simulation studies to assess statistical properties of results from these approaches.

Through assessing empirical data analysis and 18 plasmode simulation scenarios, we found that results from approaches utilizing empirical covariates are generally similar to each other and the magnitude of difference in results of these approaches are generally small. This finding is consistent with the relevant literature[6,31]. When bias-based ranking is utilized for selection of high-dimensional propensity score variables, hybrid approaches performed slightly better than the other approaches when comparing in terms of MSE, irrespective of whether unmeasured confounding was present or not. When compared with respect to bias, both high-dimensional propensity score and machine-learning approaches performed well. Exposure-based analysis results were slightly inferior to the bias-based analysis in our context, but pure machine-learning methods always performed well in these scenarios. To answer the question in the title in this paper, we were able to train the pure machine-learning methods to perform almost as good as the high-dimensional propensity score methods in many scenarios, if not better. We get even more powerful performance when we combine both approaches.

We need to consider a major limitation of this high-dimensional propensity score algo-

rithm. In the bias-based analysis, the ranking of the empirical covariates in high-dimensional propensity score analysis is done separately based on bivariate associations of the confounder and the outcome[43]. In high-dimensional setting, one can think a scenario, where many covariates are correlated, and they may contribute the same information. Thus, some of these selected high-dimensional propensity score variables might not have a confounding influence in the presence of the others[44]. Further generic limitations of this approach are outlined in eAppendix A.11. To reduce overfitting problem further, contrary to the regular high-dimensional propensity score algorithm (that considers bivariate association of the outcome and an empirical covariate), machine-learning approaches jointly consider all the empirical covariates in one multivariate model. These multivariate models follow the same epidemiologic principle that the empirical covariates associated with the outcome need to be included in the propensity score model[12].

All machine-learning methods, however, do not share the same strengths and limitations. One of the known limits with LASSO is that for a highly correlated group of variables, LASSO tends to select only one variable from a group and ignores the rest of them[29]. However, one correlated group could include more than one important confounder, and picking just one of them could potentially result in residual confounding. Elastic net is a compromise between LASSO and ridge regression and therefore, inherently more stable than a LASSO, even in the presence of severe multicollinearity. Elastic net allows selection of more than one variable from a correlated group if they are deemed sufficiently important. In terms of identifying important risk factors, data analysis examples and simulation studies have shown that the elastic net approach often outperforms the LASSO approach[28]. With high-dimensional propensity score selection as well as random forest approach, we generally do not know how many of the covariates are optimal to adjust, and generally between $200-500$ variables are considered based on subjective judgement. LASSO and elastic net select a necessary number of risk factors based on association with the outcome, and users do not have to decide how many variables to use. The computational burden associated with the machine-learning method is a cause for concern[21,22]. In high-dimensional setting, the associated computational time may be formidably high.

A number of recent studies showed that compared with a mere propensity score adjustment (using investigator-specified confounders only), further adjustment using the high-

dimensional propensity score algorithm had little or no impact on the estimates[45,46]. The propensity score building models used in these high-dimensional propensity score algorithms, such as parametric logistic regression model, are mostly historical artifacts and likely inadequate to exploit the wealth of high-dimensional administrative data properly[40,47]. In our work, in terms of MSE, we showed that the hybrid approaches, such as Hybrid-Enet and Hybrid-LASSO, that further refined the confounder selection from a chosen high-dimensional propensity score selected variable pool, performed better than the regular high-dimensional propensity score approaches in most settings.

Our findings in this paper have important implications. In all of the scenarios we have considered in this work, machine-learning and hybrid methods were shown to perform as well as or better than the conventional high-dimensional propensity score method and hence, can be considered as reliable alternatives. Routines for these machine-learning approaches are widely available in almost all of the major software packages (see eAppendix A.6) and they are easy to implement in situations where an extensive list of features (thousands of variables) are available[32]. By design, as empirical covariates are binary variables, we do not need to worry about nonlinearity while implementing the high-dimensional propensity score algorithm[48]. Also, inclusion of interactions in the high-dimensional setting generally does not affect the effect estimates much[10]. However, in the process of categorization and not assessing interactions, we do lose information that could be otherwise useful in detecting more signals from the original non-binary proxy variables using data-adaptive machine-learning algorithms. However, since the high-dimensional propensity score algorithm is dependent on Bross's formula[14], the current high-dimensional propensity score algorithm is constrained only to handle binary covariates, binary exposure and binary outcomes. Regarding handling various types of variables, many of the pure machine-learning methods are free from such limitation in general and can be easily extended to handle continuous, count or survival outcomes[49,50] as well as various types of covariates (binary, count, continuous).

The high-dimensional propensity score based analyses are done based on a strong assumption that the selected empirical covariates collectively serve as proxies for all unmeasured or residual confounders[44]. As a result of this assumption, residual confounding is thought to be adjusted by high-dimensional propensity score analysis. However, this assumption is not empirically verifiable and hence debatable. When we use high-dimensional propensity

score or alternative machine-learning methods, we do expect to reduce the effect of residual confounding to some extent, but eliminating residual confounding completely is unlikely in a real-life setting. The scope of the bias reduction will generally depend on the availability of the right surrogates of the unmeasured or imperfectly observed factors[9,10]. As seen in the sensitivity analysis from our empirical data analysis example, none of the methods adjusting for numerous proxy variables were able to compensate for the omitted confounders fully. As a general rule of thumb, one should always consider doing a regular propensity score analysis first and then perform a high-dimensional propensity score analysis. That way, one can have a sense of the amount and direction of correction and adjustment. In our simulations, high-dimensional propensity score, machine-learning methods, and hybrid approaches utilizing the empirical covariates always performed better than a regular propensity score analysis.

References

[1] M. E. Karim, P. Gustafson, J. Petkau, Y. Zhao, A. Shirani, E. Kingwell, C. Evans, M. van der Kop, J. Oger, and H. Tremlett. Marginal Structural Cox Models for Estimating the Association Between β-Interferon Exposure and Disease Progression in a Multiple Sclerosis Cohort. *American Journal of Epidemiology*, 180(2):160–171, 2014.

[2] M Alan Brookhart, Richard Wyss, J Bradley Layton, and Til Stürmer. Propensity score methods for confounding control in nonexperimental research. *Circulation: Cardiovascular Quality and Outcomes*, 6(5):604–611, 2013.

[3] M.E. Karim. Can joint replacement reduce cardiovascular risk? *British Medical Journal*, 347:f6651, 2013.

[4] B.A. Brumback, M.A. Hernán, S.J.P.A. Haneuse, and J.M. Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5):749–767, 2004.

[5] Lawrence C McCandless and Paul Gustafson. A comparison of bayesian and monte carlo sensitivity analysis for unmeasured confounding. *Statistics in Medicine*, 2017. DOI:10.1002/sim.7298.

[6] 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.

[7] Til Stürmer, Sebastian Schneeweiss, Jerry Avorn, and Robert J Glynn. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology*, 162(3):279–289, 2005.

[8] M.E. Karim and Paul Gustafson. Hypothesis testing for an exposure–disease association in case–control studies under nondifferential exposure misclassification in the presence of validation data: Bayesian and frequentist adjustments. *Statistics in Biosciences*, 2(8):234–252, 2016.

[9] Sander Greenland. The effect of misclassification in the presence of covariates. *American journal of epidemiology*, 112(4):564–569, 1980.

[10] 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.

[11] Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

[12] M.A. Brookhart, S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[13] 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.

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

[15] J.A. Rassen, R.J. Glynn, M.A. Brookhart, and S. Schneeweiss. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American journal of epidemiology*, 173(12):1404–1413, 2011.

[16] 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.

[17] Sander Greenland. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167(5):523–529, 2008.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[19] S. Rose. Mortality risk score prediction in an elderly population using machine learning. *American journal of epidemiology*, 177(5):443–452, 2013.

[20] R. Pirracchio, M.L. Petersen, and M. van der Laan. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*, 181(2):108–119, 2015.

[21] M.E. Karim, J. Petkau, P. Gustafson, and H. Tremlett. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: Hedging against weight-model misspecification, 2016. DOI: 10.1080/03610918.2016.1248574, published online: 21 Oct 2016.

[22] M.E. Karim and R.W. Platt. Estimating inverse probability weights using super learner

when weight-model specification is unknown in a marginal structural cox model context. *Statistics in Medicine*, 36(13):2032–2047, 2017.

[23] M.E. Karim. *Causal inference approaches for dealing with time-dependent confounding in longitudinal studies, with applications to multiple sclerosis research*. PhD thesis, University of British Columbia, 2015.

[24] J.M. Franklin, W.H. Shrank, J. Lii, A.K. Krumme, O.S. Matlin, T.A. Brennan, and N.K. Choudhry. Observing versus predicting: Initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. *Health services research*, 51(1):220–239, 2016.

[25] J. Pearl. Invited commentary: understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227, 2011.

[26] 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.

[27] 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.

[28] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[29] Trevor Hastie and Junyang Qian. Glmnet vignette, 2014.

[30] Leo Breiman. Random forests, 1999.

[31] 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.

[32] 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.

[33] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[34] C. Ju, M. Combs, S.D. Lendle, J.M. Franklin, R. Wyss, S. Schneeweiss, and M.J. van der Laan. Propensity score prediction for electronic healthcare dataset using super learner

and high-dimensional propensity score method. Website, 2016. Last accessed: July 27. URL: http://biostats.bepress.com/ucbbiostat/paper351/.

[35] J.S. Kaufman. Marginalia: comparing adjusted effect measures. *Epidemiology*, 21(4):490–493, 2010.

[36] M.E. Karim, J. Petkau, P. Gustafson, R.W. Platt, and H. Tremlett. Comparison of statistical approaches dealing with time-dependent confounding in drug effectiveness studies. *Statistical Methods in Medical Research*, page 0962280216668554, 2016. DOI: 10.1177/0962280216668554.

[37] R. Wyss, S. Schneeweiss, M. van der Laan, S.D. Lendle, C. Ju, and J.M. Franklin. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*, pages 1–33, 2017. DOI: 10.1097/EDE.0000000000000762.

[38] John M Brooks and Robert L Ohsfeldt. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health services research*, 48(4):1487–1507, 2013.

[39] S. Gruber, R.W. Logan, I. Jarrín, S. Monge, and M.A. Hernán. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in Medicine*, 2014.

[40] R. Neugebauer, J.A. Schmittdiel, Z. Zhu, J.A. Rassen, J.D. Seeger, and S. Schneeweiss. High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Statistics in medicine*, 34(5):753–781, 2014.

[41] B.K. Lee, J. Lessler, and E.A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

[42] Mark J van der Laan. Targeted maximum likelihood based causal inference: Part i. *The International Journal of Biostatistics*, 6(2), 2010.

[43] J.A. Rassen and S. Schneeweiss. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and drug safety*, 21(S1):41–49, 2012.

[44] Dirk Enders, Christoph Ohlmeier, and Edeltraut Garbe. The potential of high-dimensional propensity scores in health services research: An exemplary study on the quality of care for elective percutaneous coronary interventions. *Health Services Research*, pages 1–17, 2017.

[45] J.R. Guertin, E. Rahme, C.R. Dormuth, and J. LeLorier. Head to head comparison

of the propensity score and the high-dimensional propensity score matching methods. *BMC medical research methodology*, 16(1):1–10, 2016.

[46] S. Toh, L.A. García Rodríguez, and M.A. Hernán. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiology and drug safety*, 20(8):849–857, 2011.

[47] S. Gruber, R.W. Logan, I. Jarrín, S. Monge, and M.A. Hernán. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in medicine*, 34(1):106–117, 2015.

[48] Krista F Huybrechts, M Alan Brookhart, Kenneth J Rothman, Rebecca A Silliman, Tobias Gerhard, Stephen Crystal, and Sebastian Schneeweiss. Comparison of different approaches to confounding adjustment in a study on the association of antipsychotic medication with mortality in older nursing home patients. *American journal of epidemiology*, 174(9):1089–1099, 2011.

[49] Robert Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

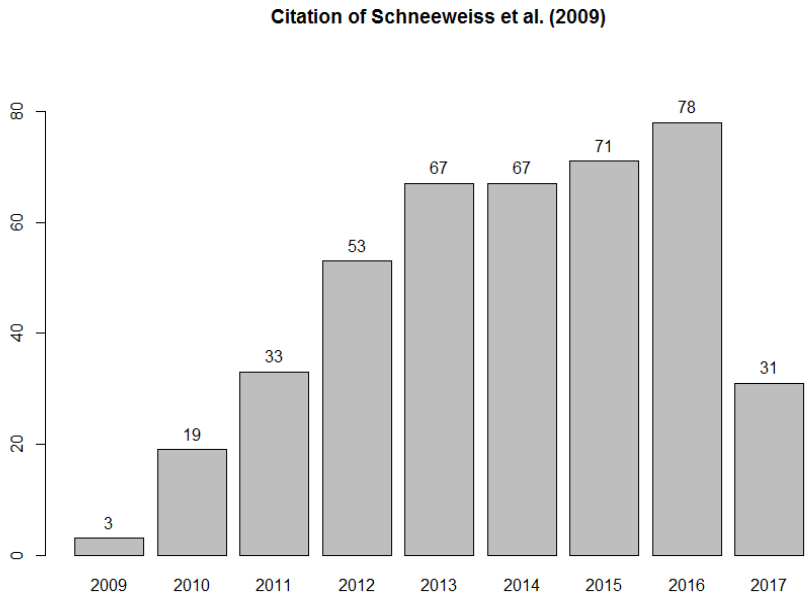
[50] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

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*

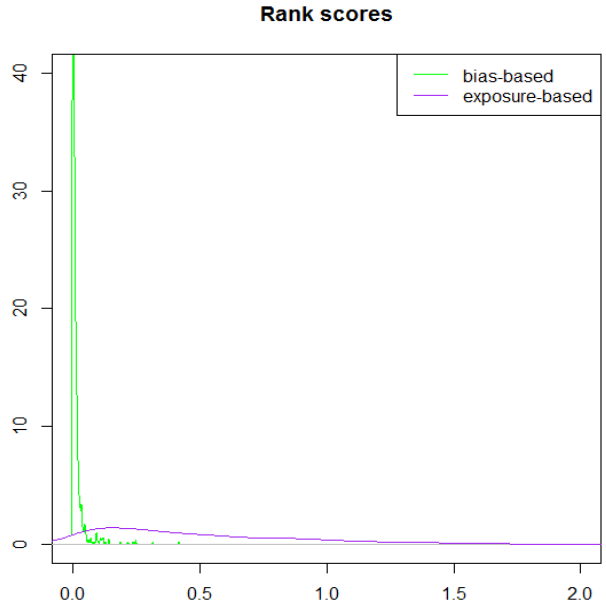
eFigure A.2: Ranking by log-bias score

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

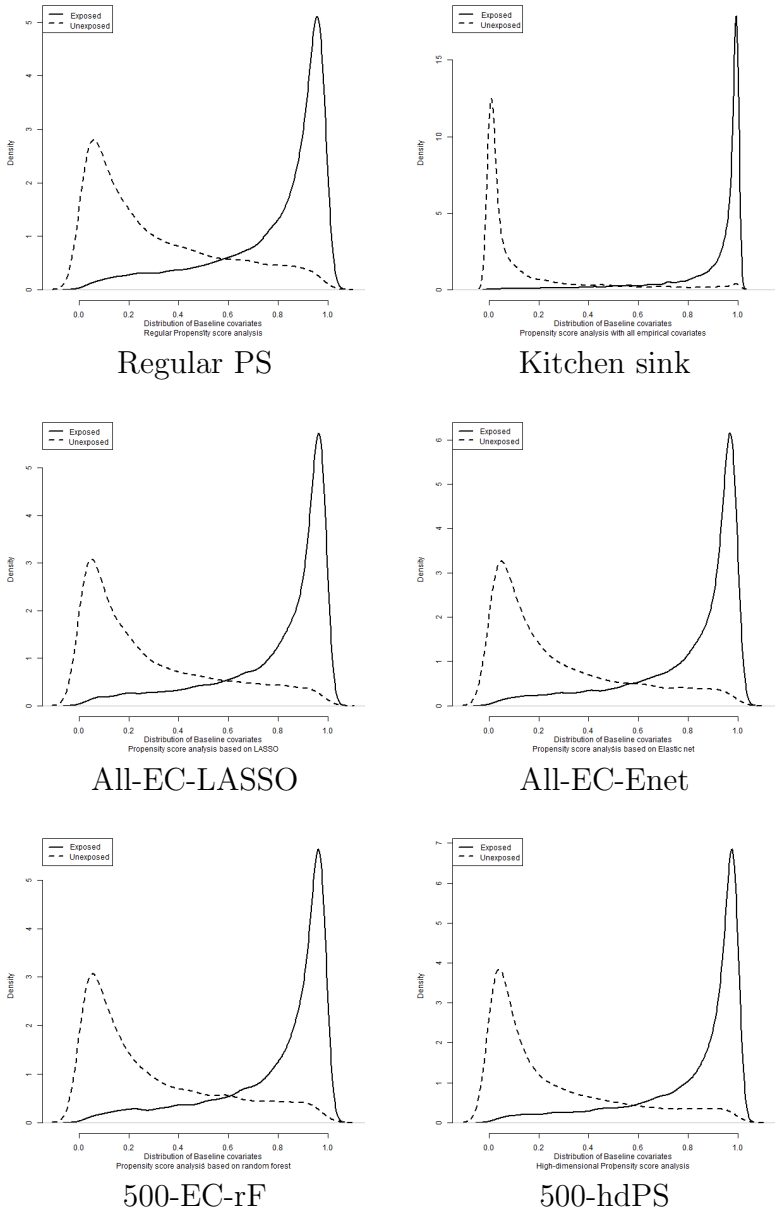
γ parameter vector respectively. In certain combination of these parameters values, it is possible that an important confounder in the empirical study may not remain important in the plasmode samples. Future research should investigate further in this issue.

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.

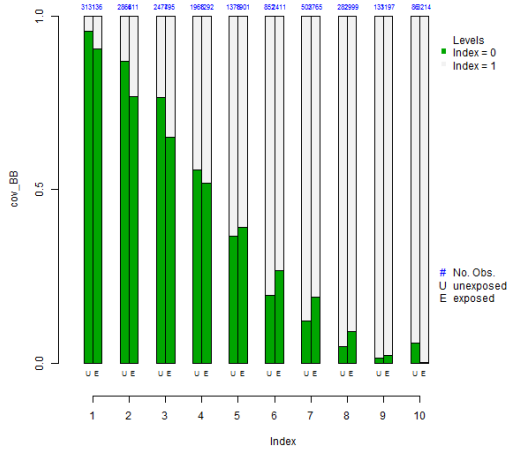
A.8 Balance diagnostics and Data Analyses

A.8.1 Balance diagnostics

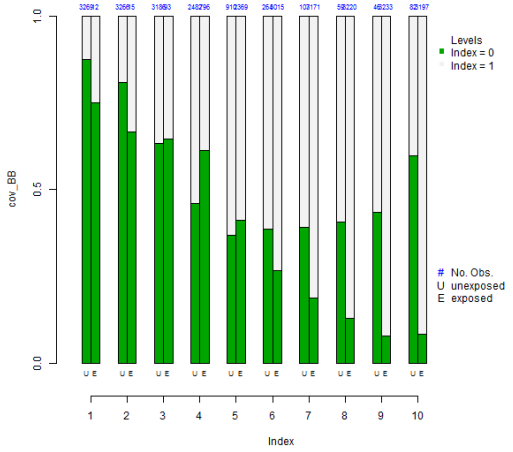


eFigure A.4: Balance

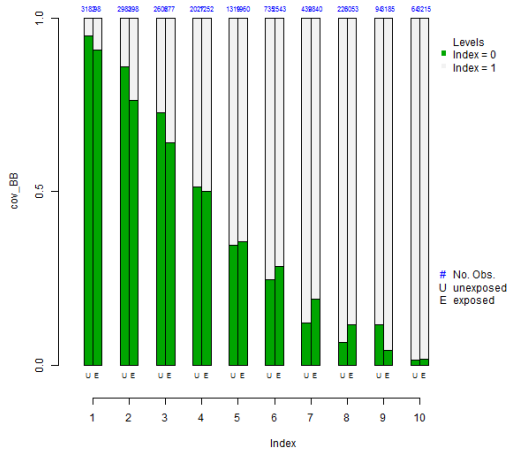
For the purpose of illustration, we checked the balance of the beta-blocker covariate, and we observe that there are imbalances in the last few deciles of PSs when we considered all empirical-covariates. However, when we selected the 500 top ranked hdPS variables, the balance is regained (see eFigure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



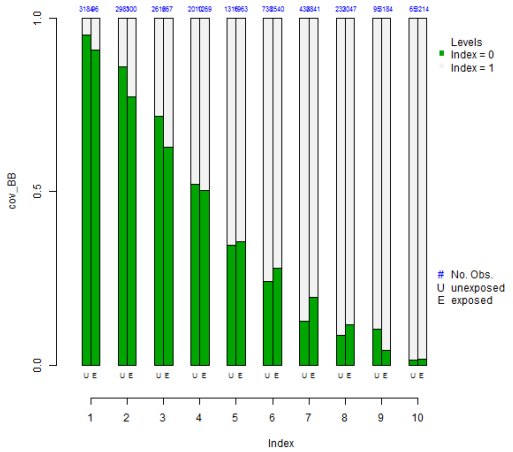
Regular PS



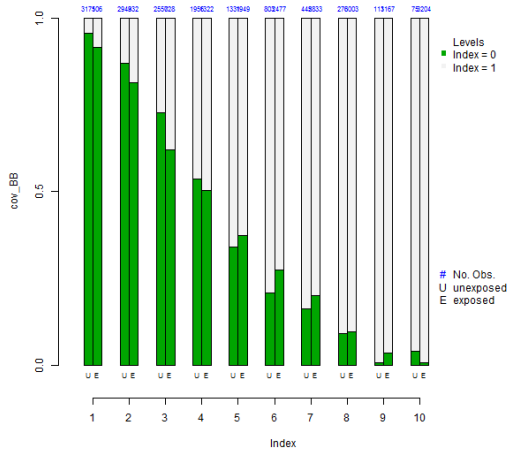
Kitchen sink



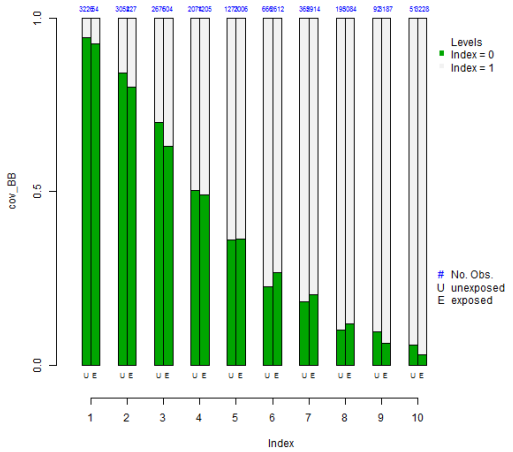
All-EC-LASSO



All-EC-Enet



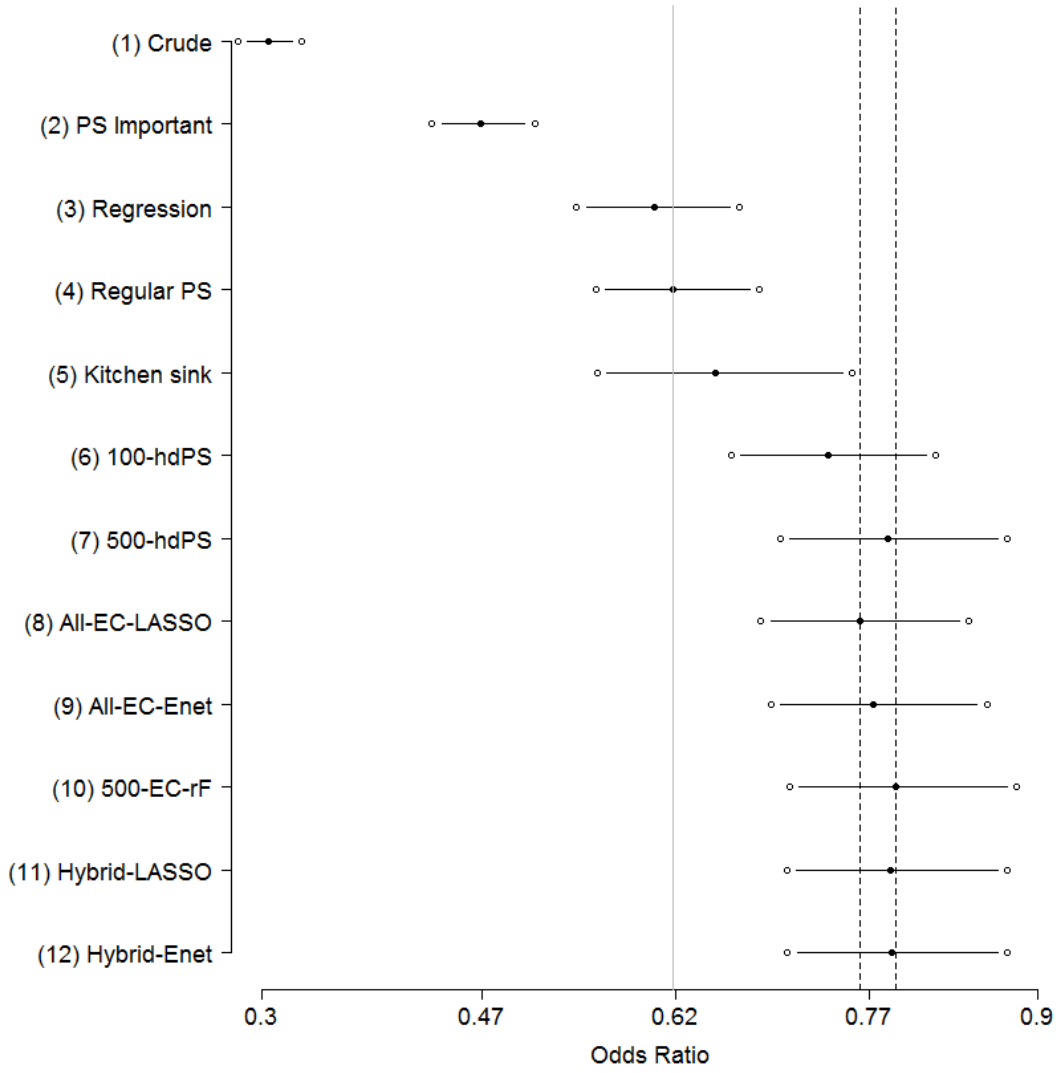
500-EC-rF



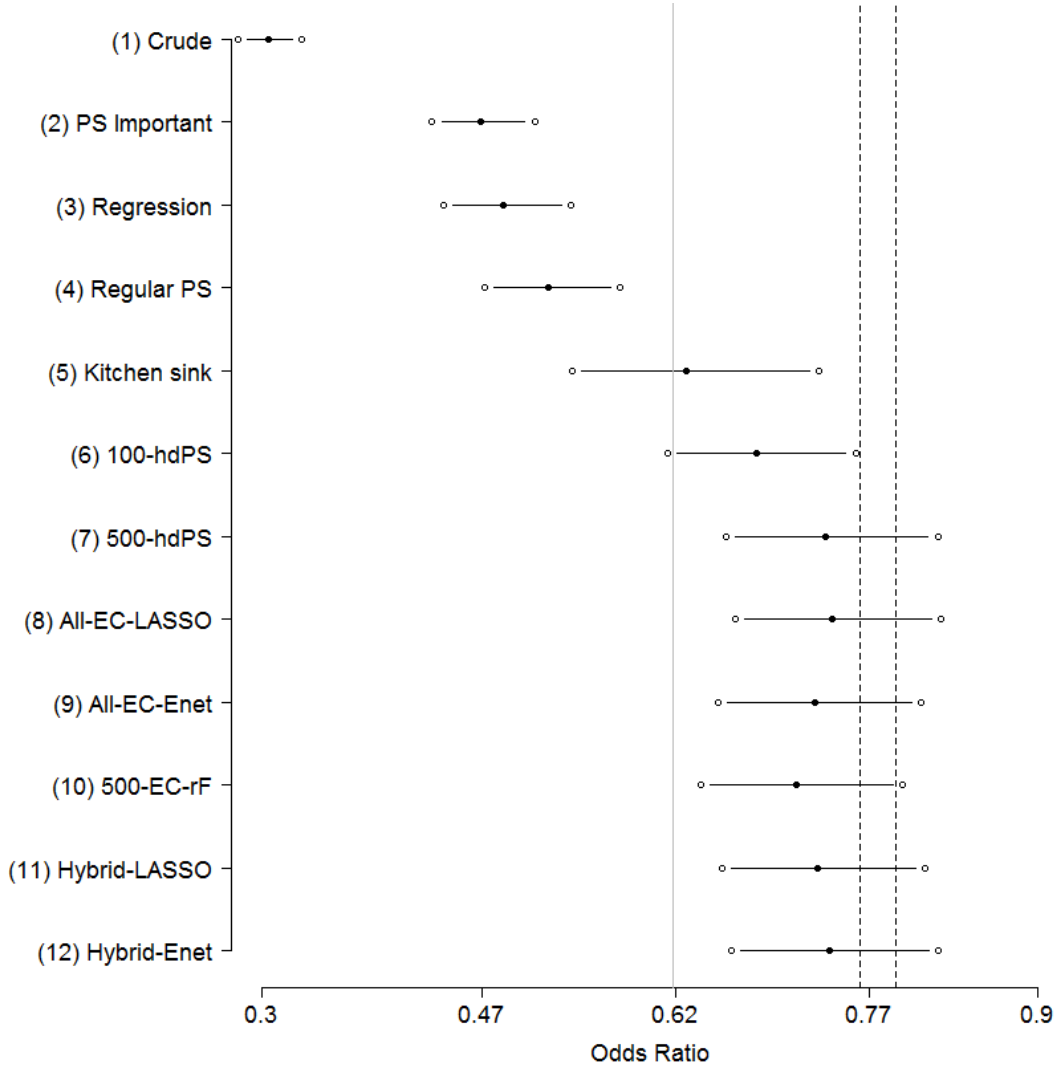
500-hdPS

eFigure A.5: Balance for beta clocker

A.8.2 Data Analyses



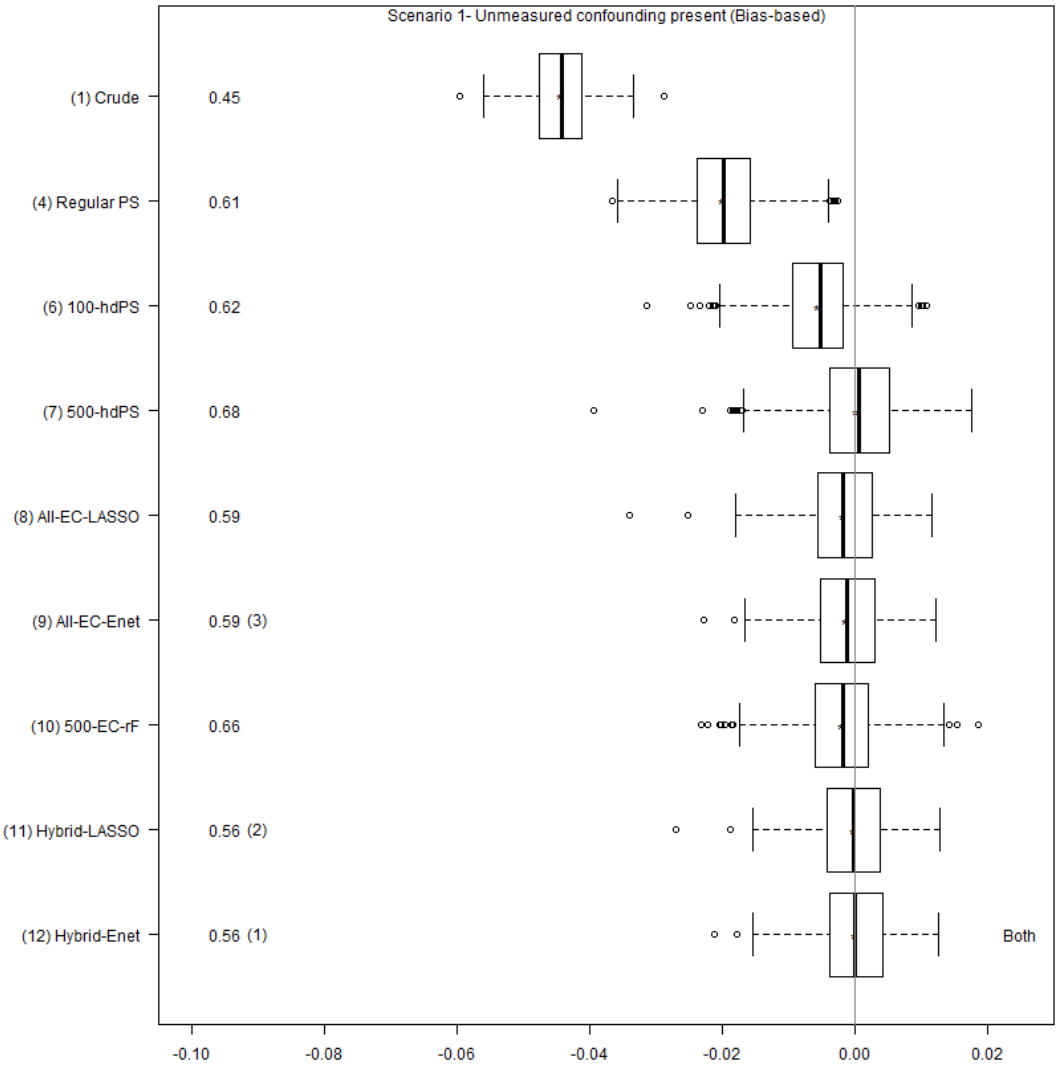
eFigure A.6: Analysis results from the approaches under consideration. When only the investigator-specified covariates were considered, the estimated OR was 0.62 in our analysis (represented by the solid grey line). When considering 500 or more empirical covariates and all the investigator-specified 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.



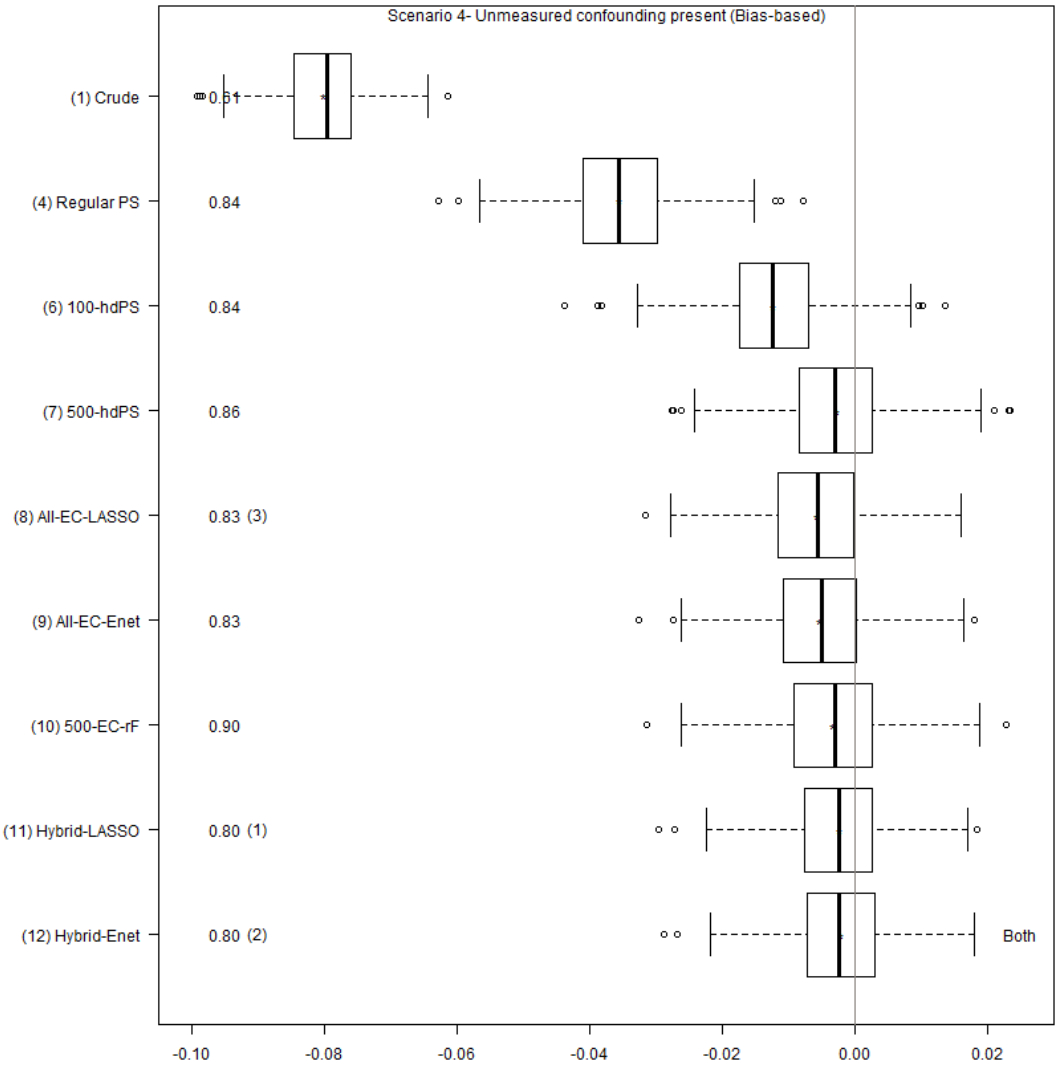
eFigure A.7: High-dimensional propensity score and machine learning alternative results without the five important covariates: age, sex, obesity, smoking, and history of diabetes. For comparison with the analyses with these five covariates, the solid grey line represents the estimated OR of 0.62 (when 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

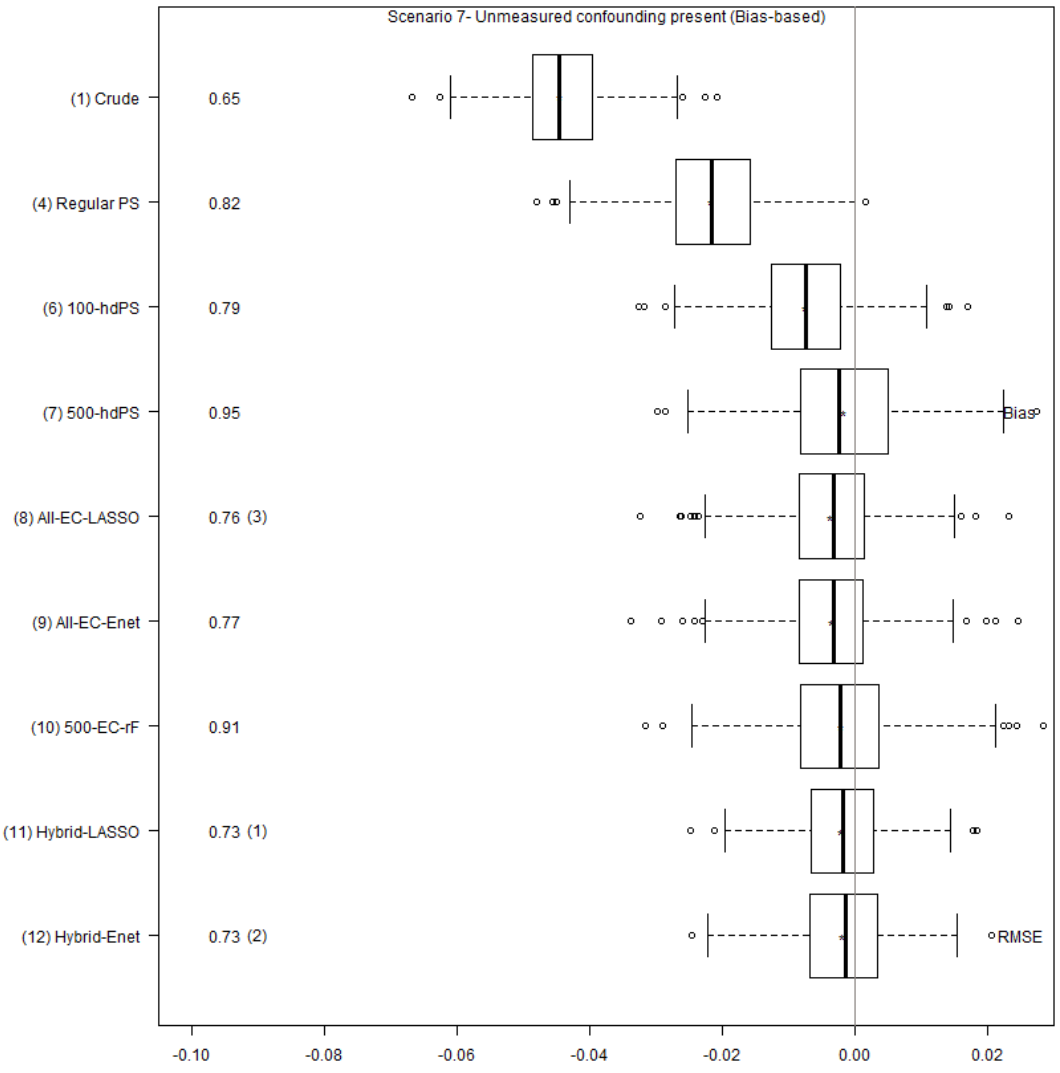
A.9.1 Unmeasured confounding present (Bias-based analysis): Main three scenarios:



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.

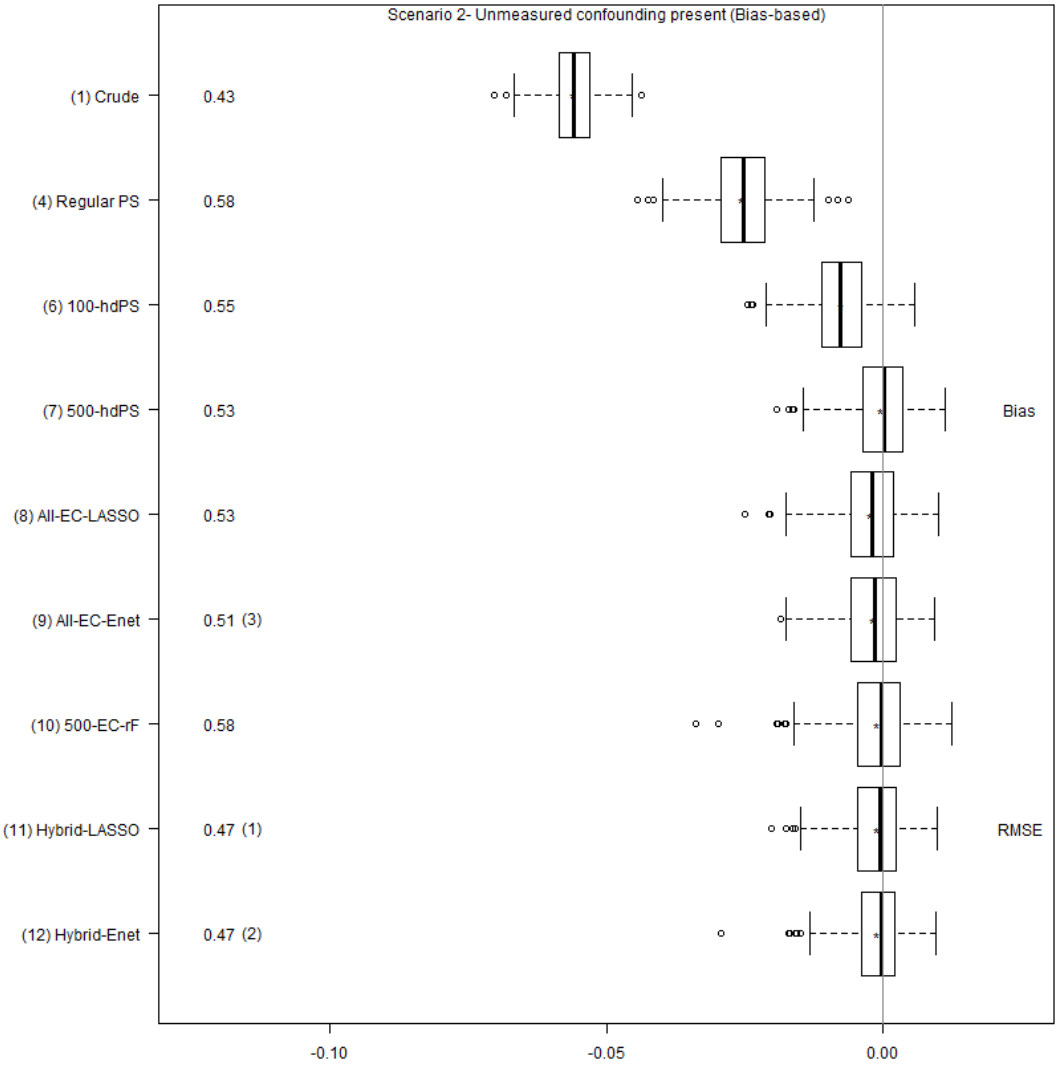


eFigure A.9: Side-by-side boxplots of the estimated risk differences (from 500 datasets) via the approaches under consideration in the plasmode Simulation Scenario 4-U. Corresponding mean values are marked by $*$. The indicator "Both" means the approach is found best by both MSE and bias criteria. Abbreviations: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

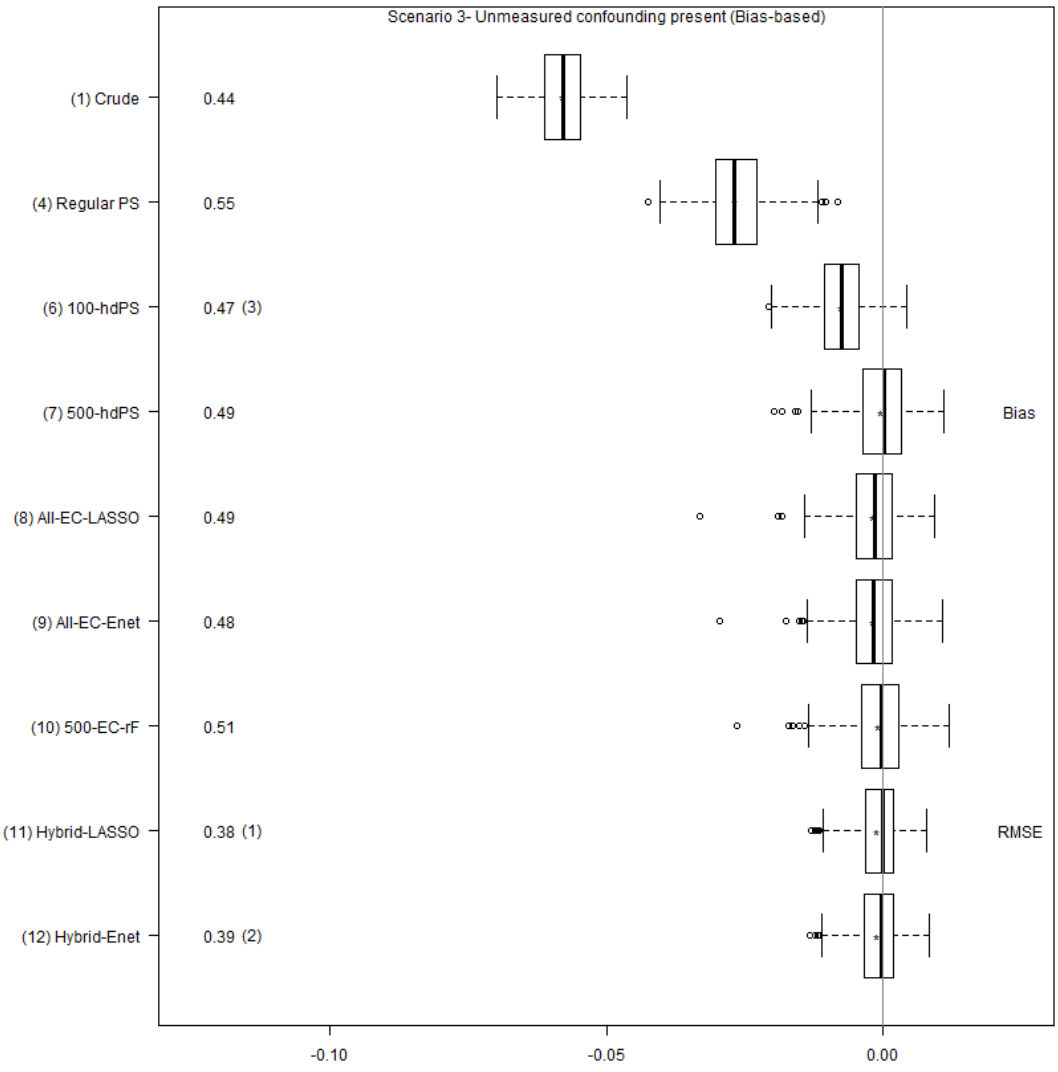


eFigure A.10: Side-by-side boxplots of the estimated risk differences (from 500 datasets) via the approaches under consideration in the plasmode Simulation Scenario 7-U. Corresponding mean values are marked by $*$. The indicator "RMSE" means the approach is found best by the RMSE criterion and the indicator "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.

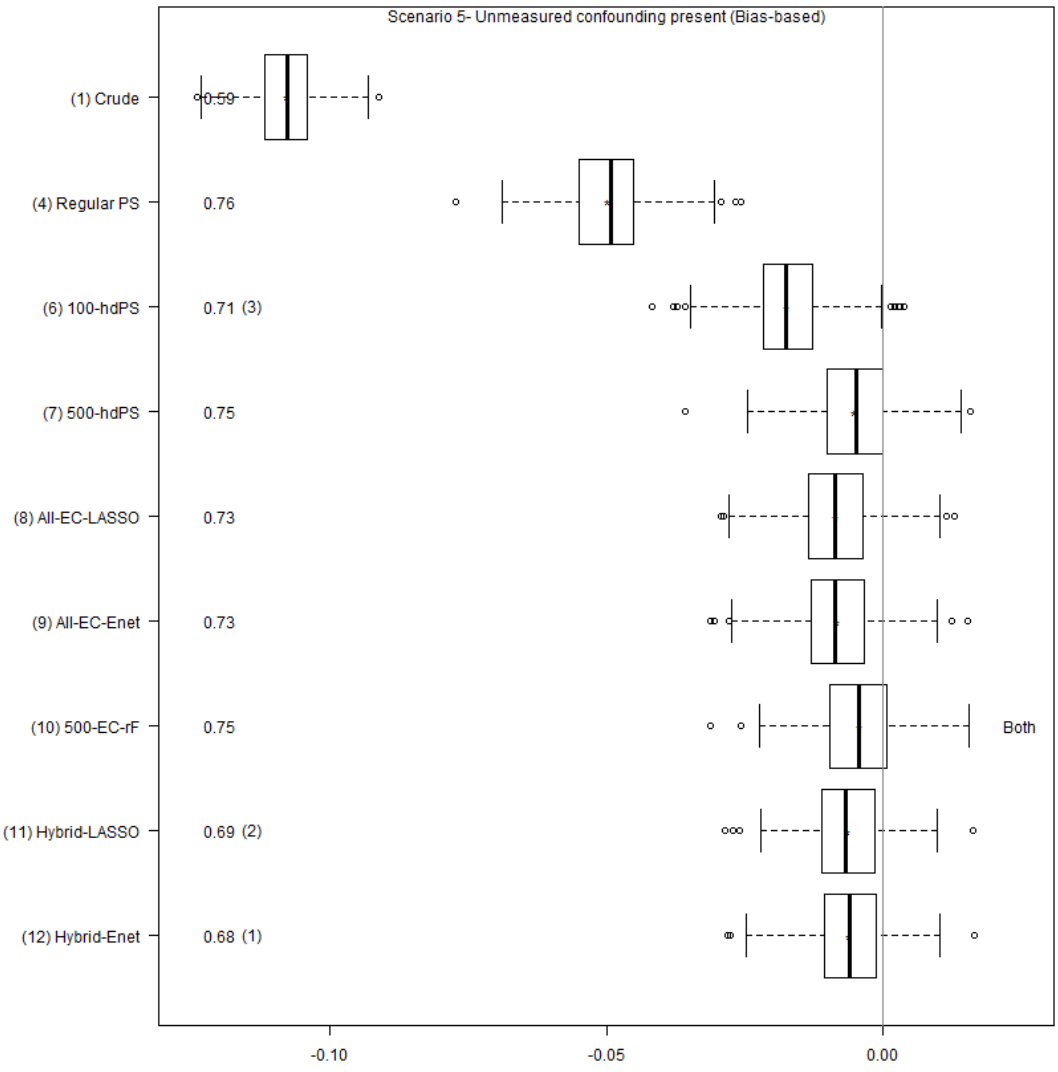
A.9.2 Unmeasured confounding present (Bias-based analysis): Other scenarios:



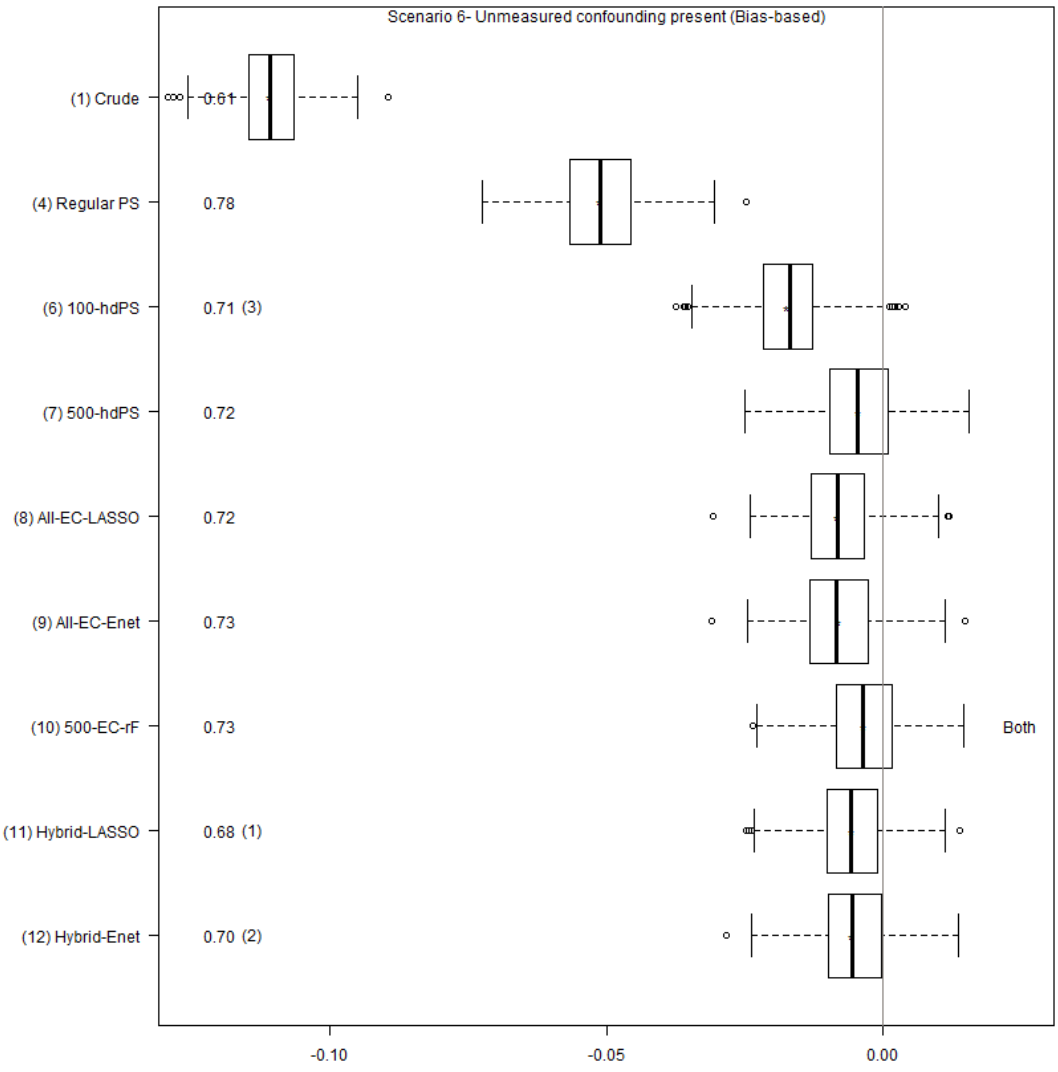
eFigure A.11: Plasmode Simulation Scenario 2-U



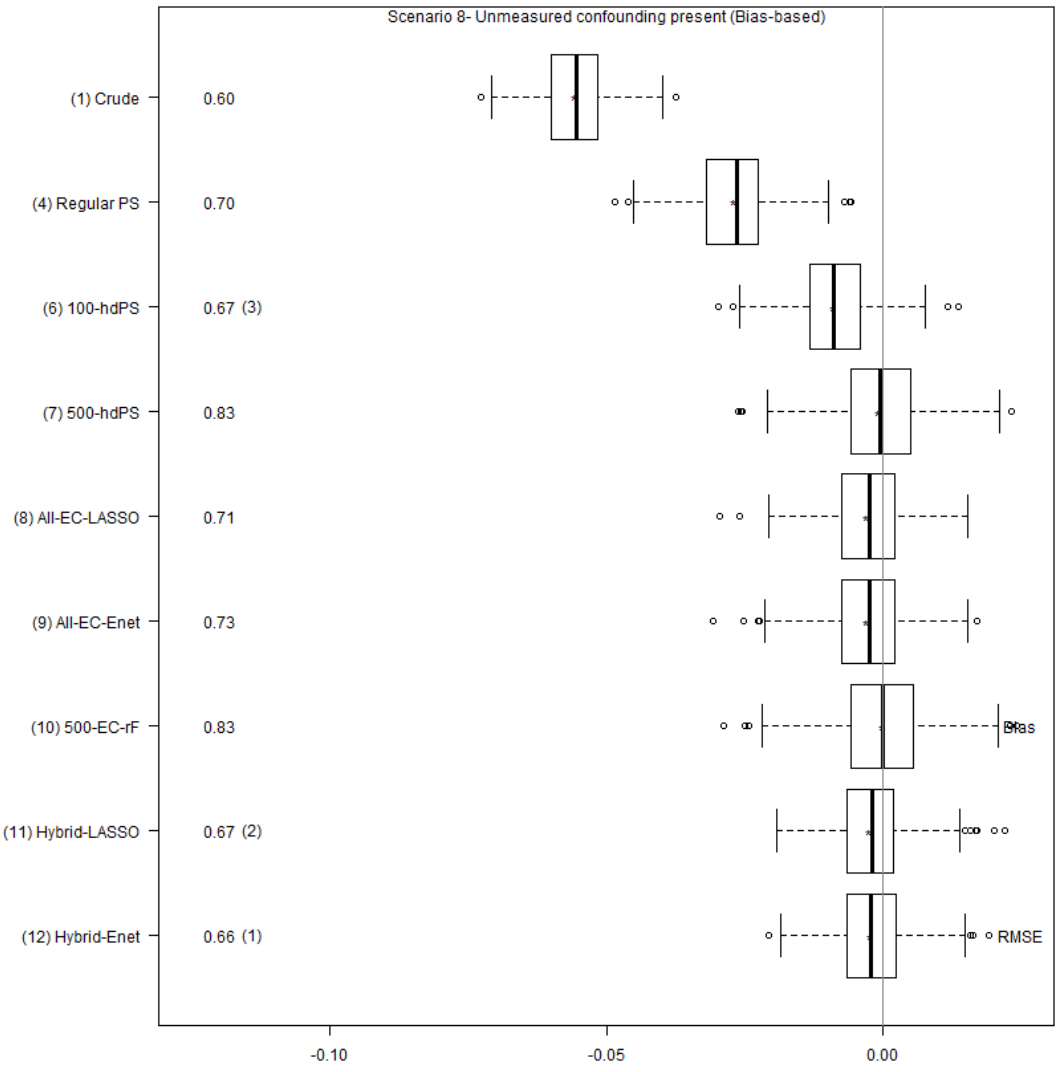
eFigure A.12: Plasmode Simulation Scenario 3-U



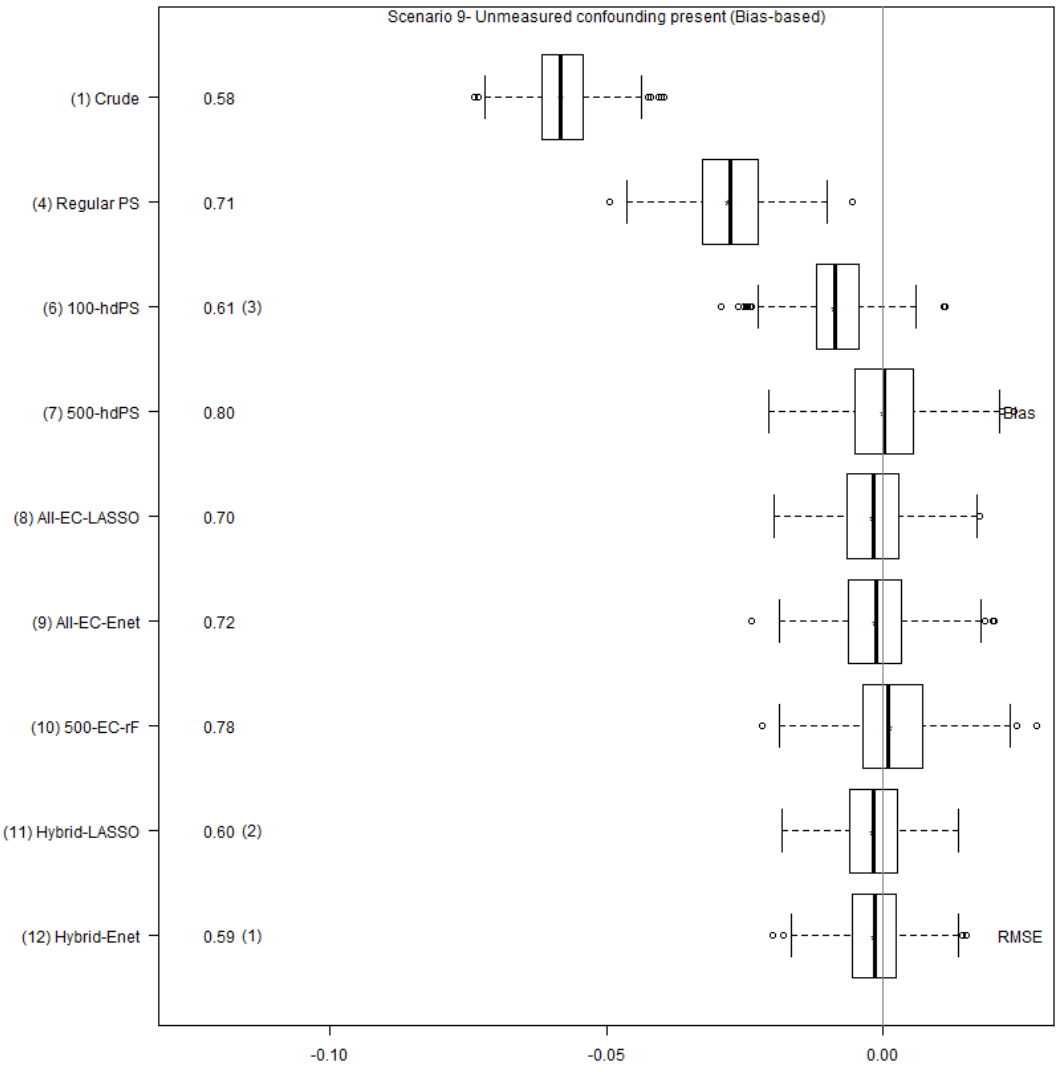
eFigure A.13: Plasmode Simulation Scenario 5-U



eFigure A.14: Plasmode Simulation Scenario 6-U

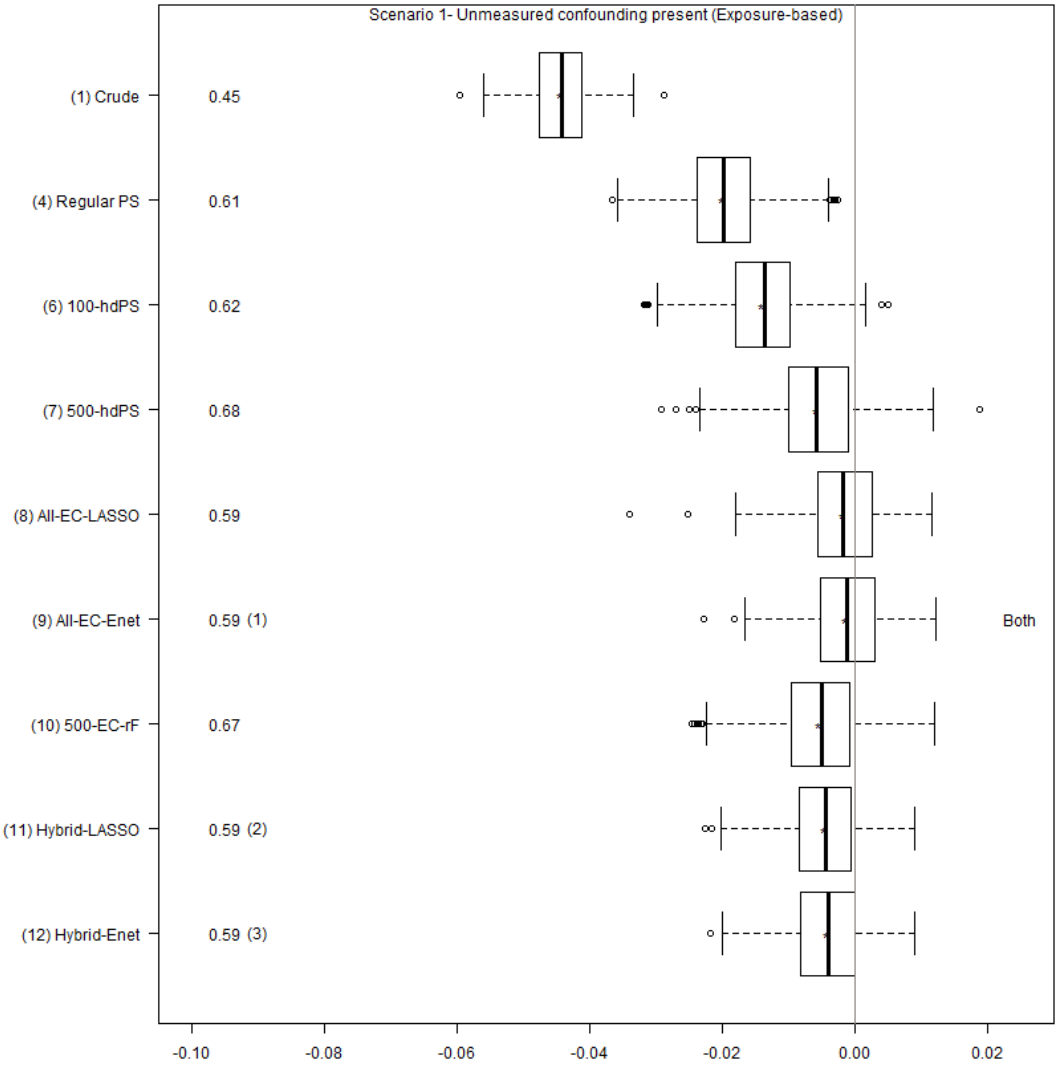


eFigure A.15: Plasmode Simulation Scenario 8-U

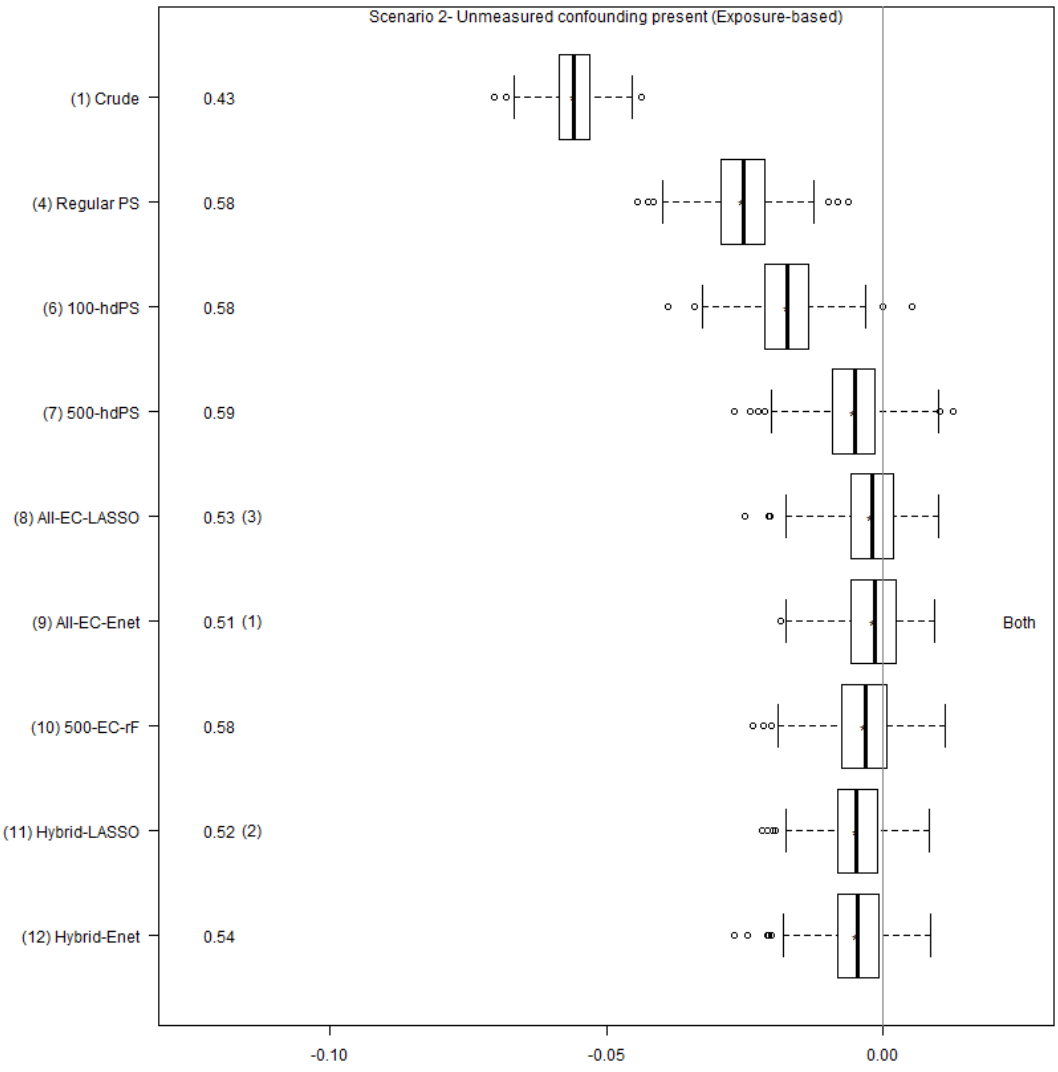


eFigure A.16: Plasmode Simulation Scenario 9-U

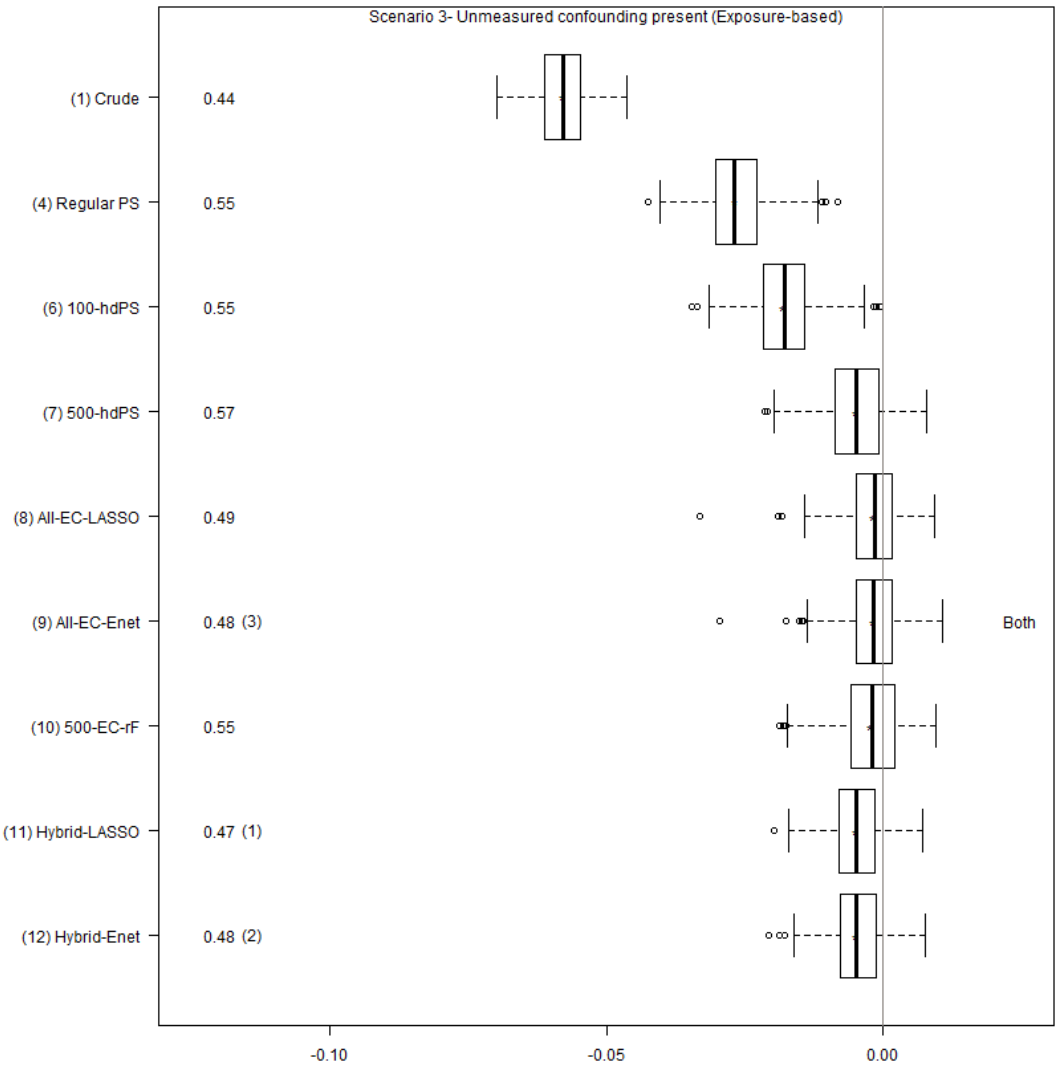
A.9.3 If unmeasured confounding present (Exposure-based analysis)



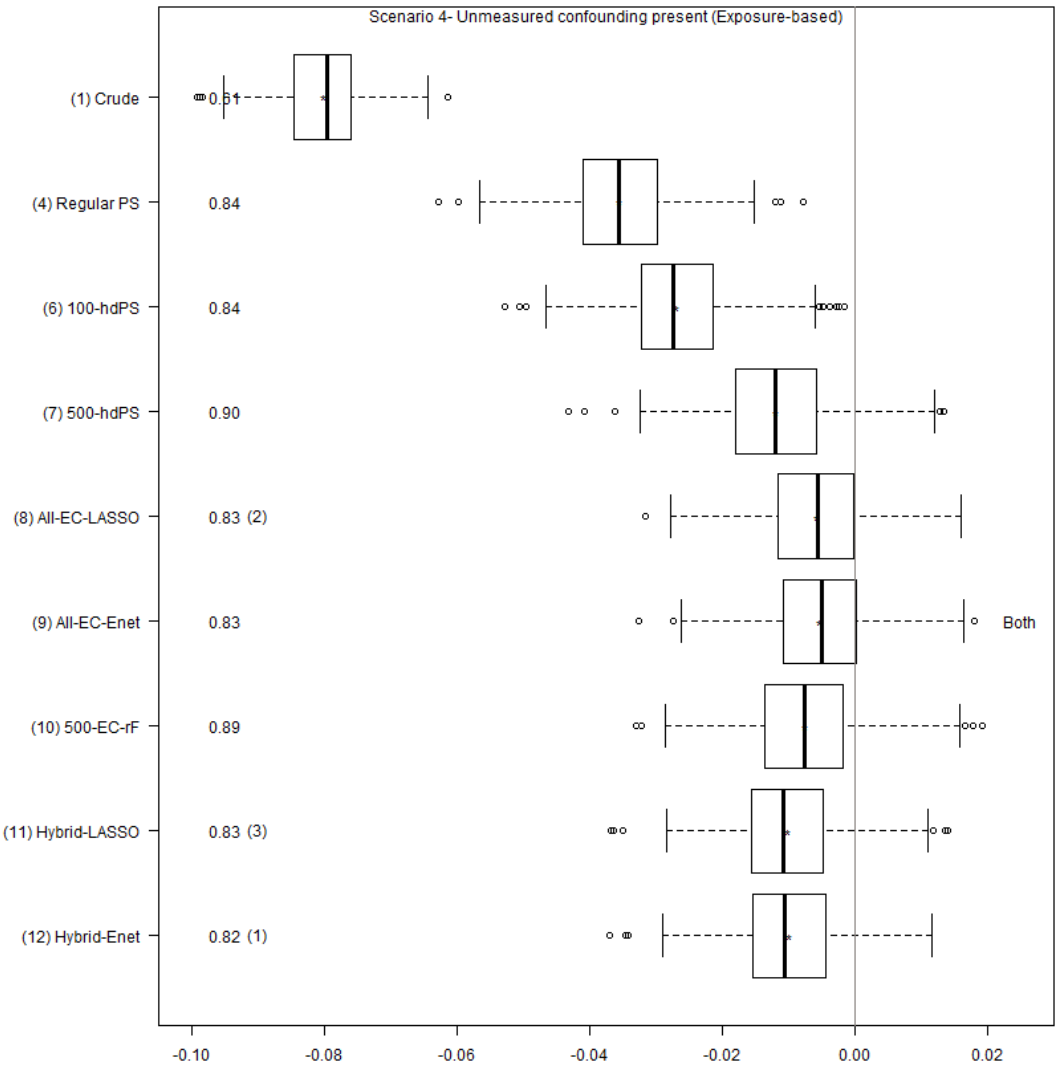
eFigure A.17: Plasmode Simulation Scenario 1-A



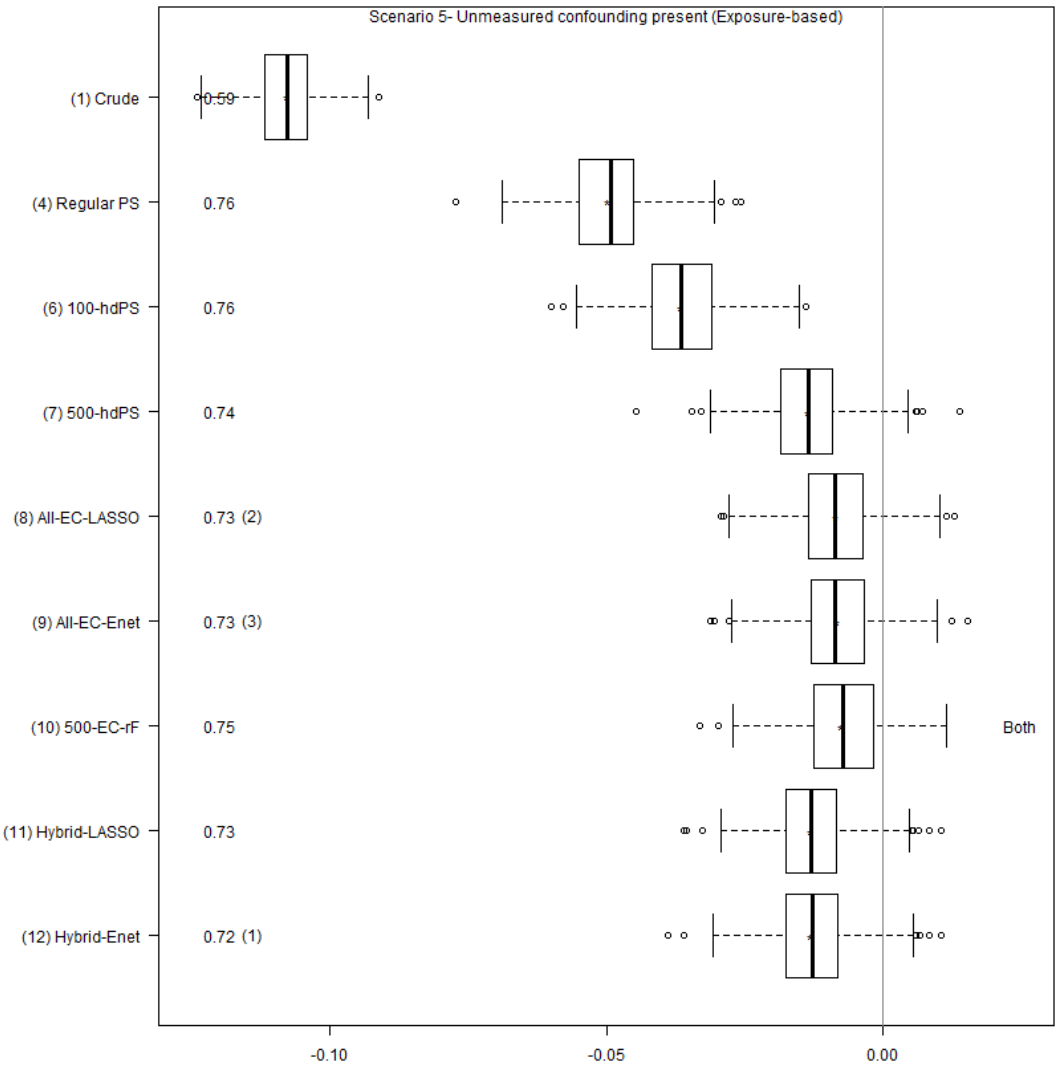
eFigure A.18: Plasmode Simulation Scenario 2-A



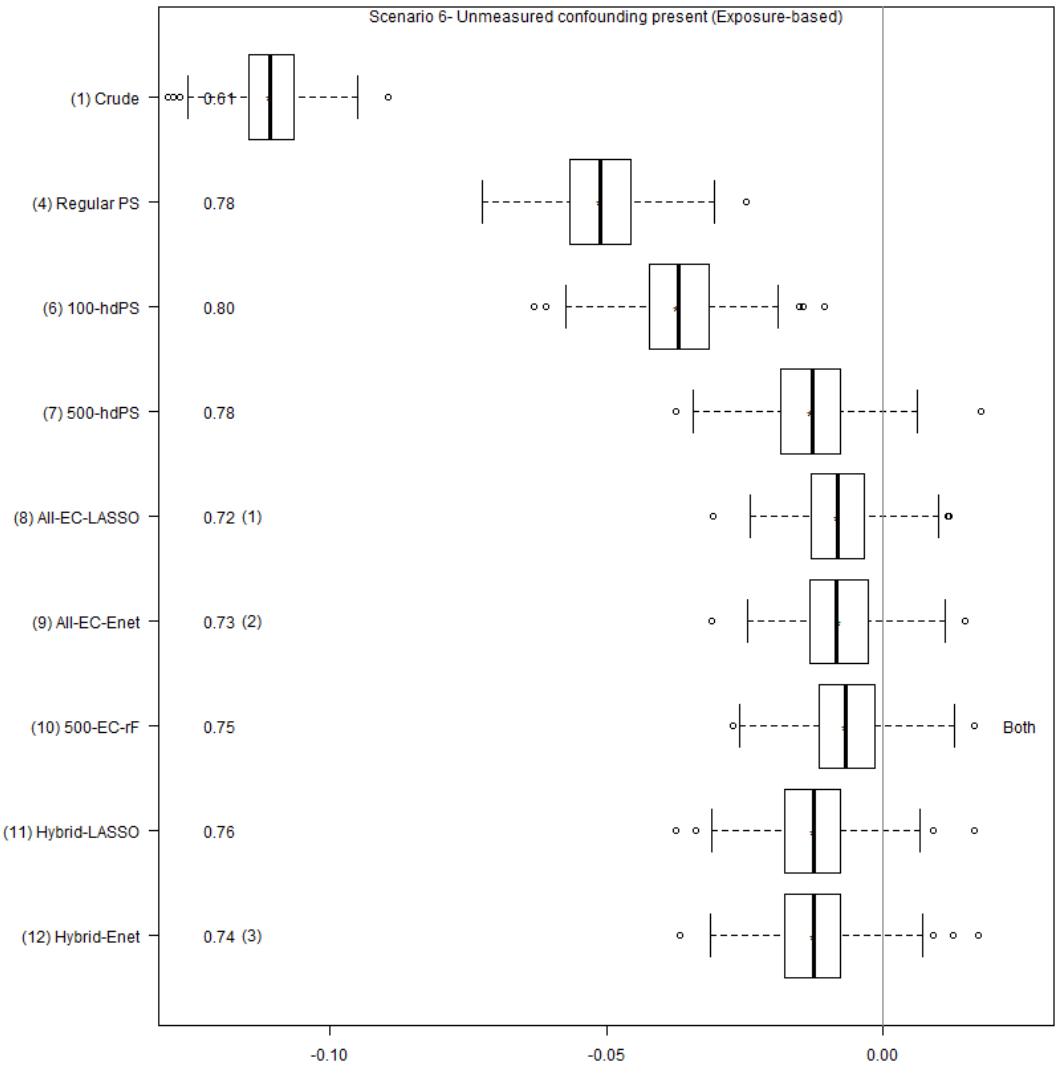
eFigure A.19: Plasmode Simulation Scenario 3-A



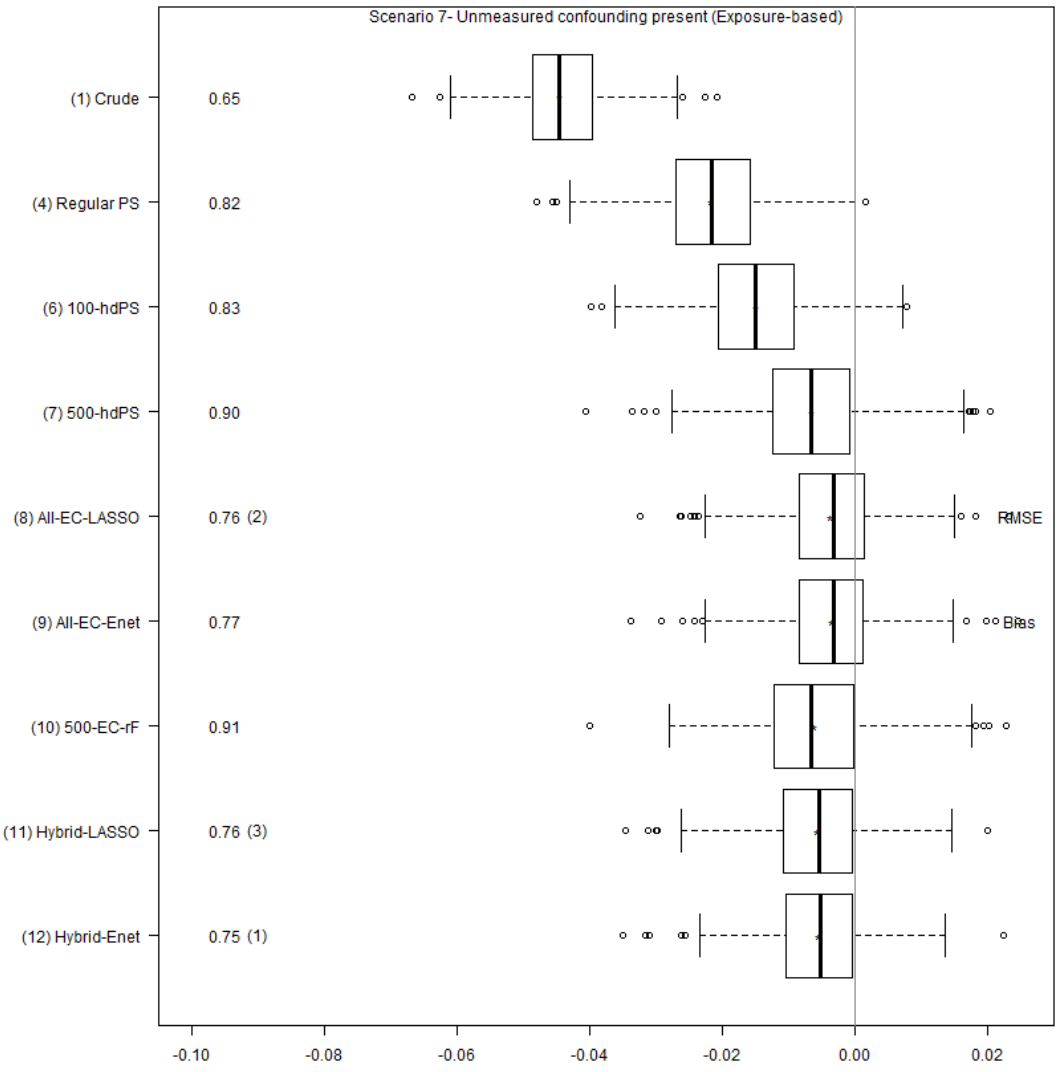
eFigure A.20: Plasmode Simulation Scenario 4-A



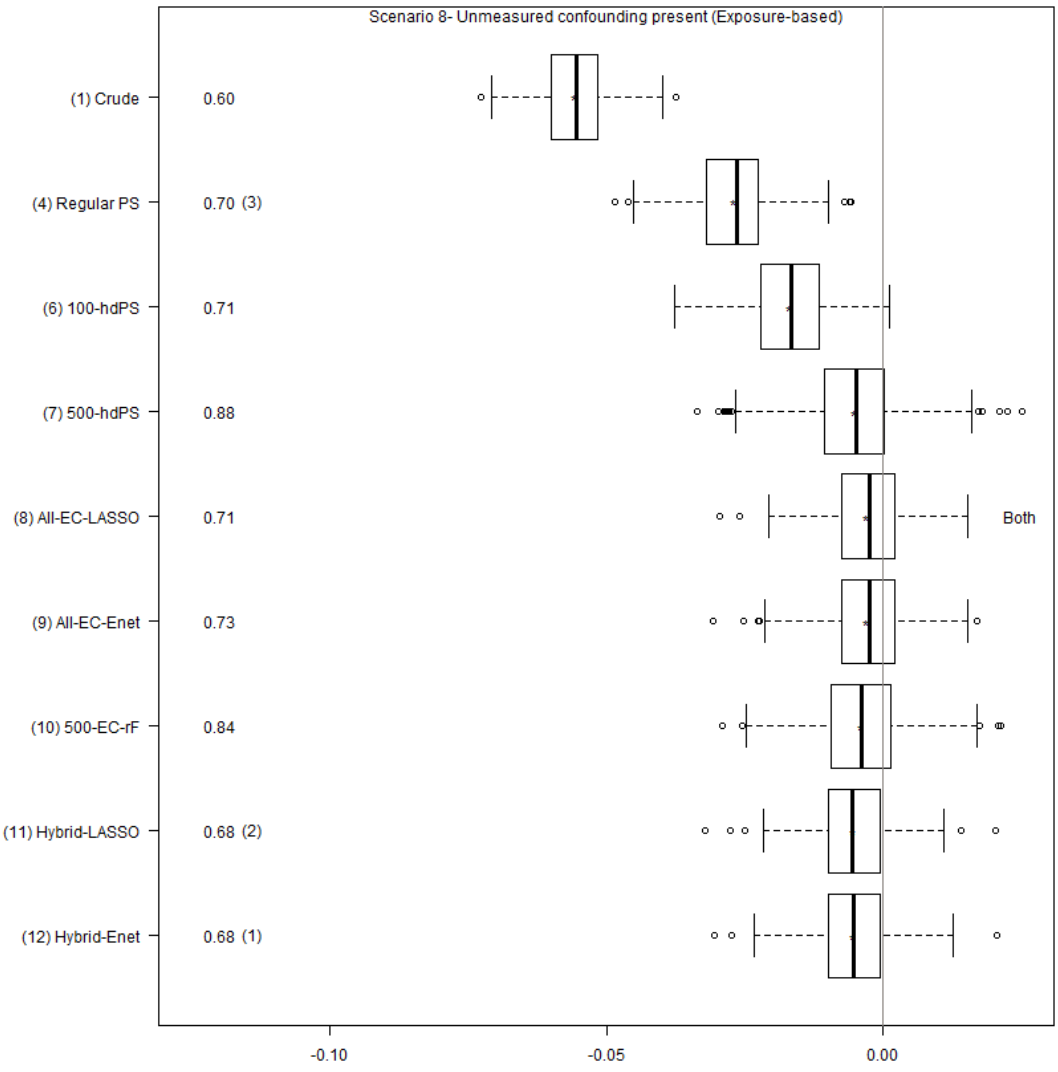
eFigure A.21: Plasmode Simulation Scenario 5-A



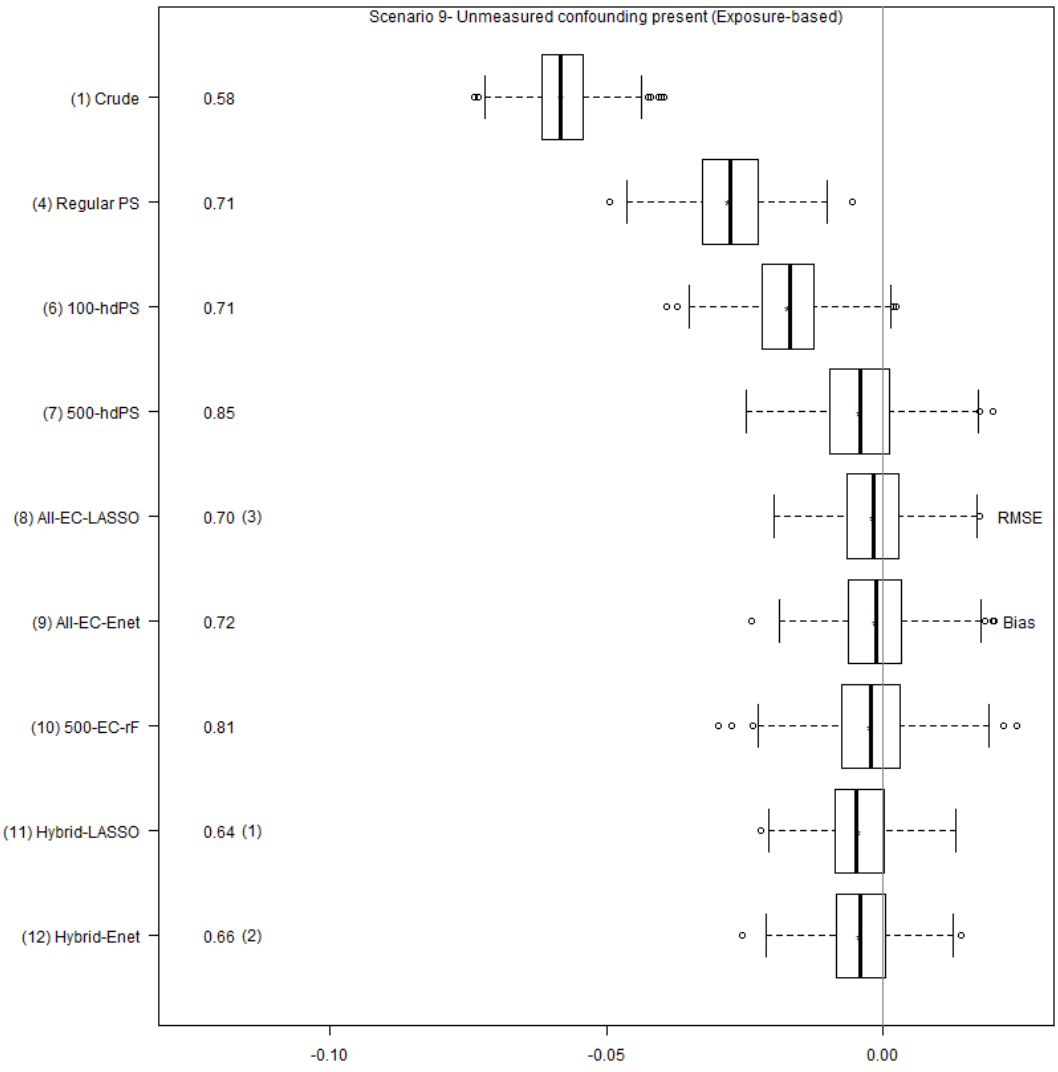
eFigure A.22: Plasmode Simulation Scenario 6-A



eFigure A.23: Plasmode Simulation Scenario 7-A

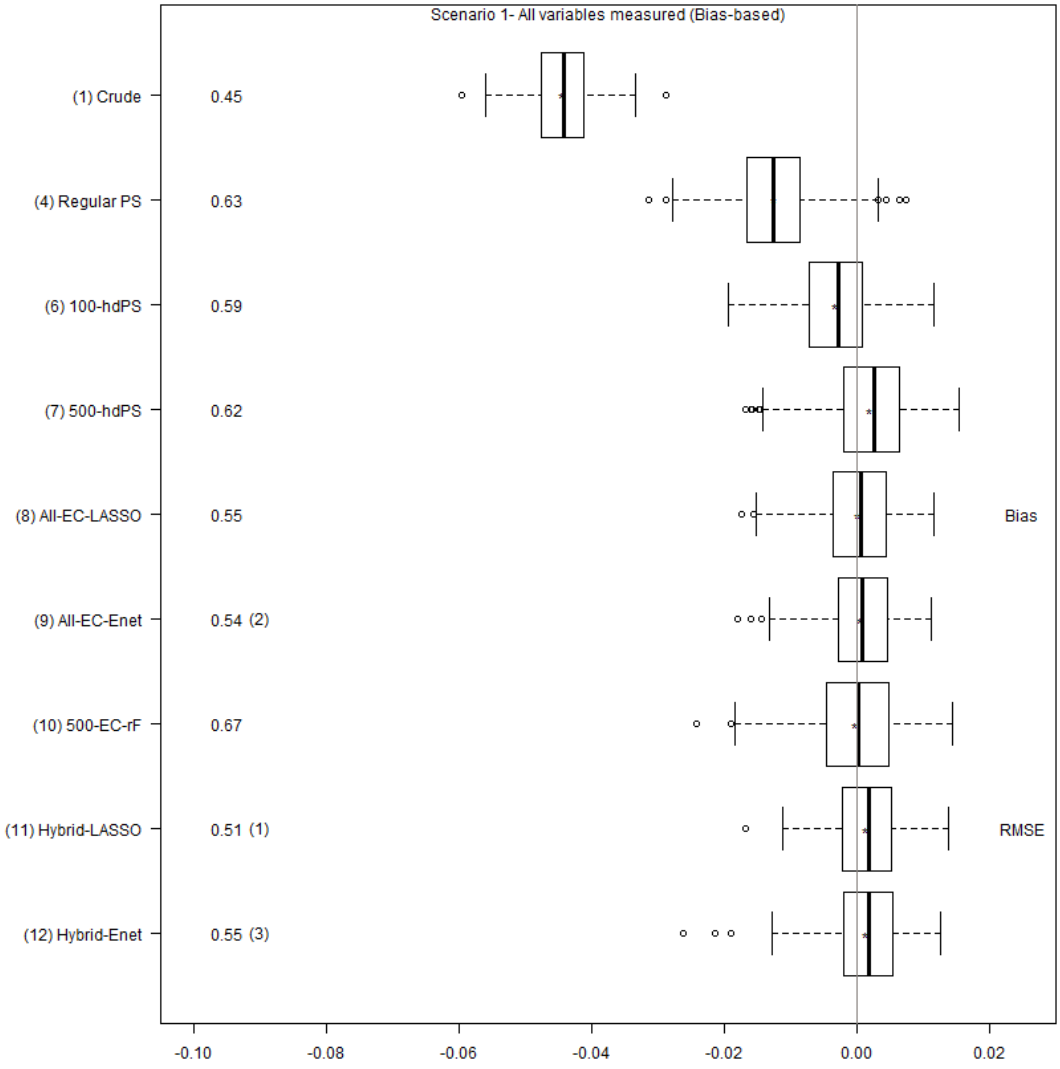


eFigure A.24: Plasmode Simulation Scenario 8-A

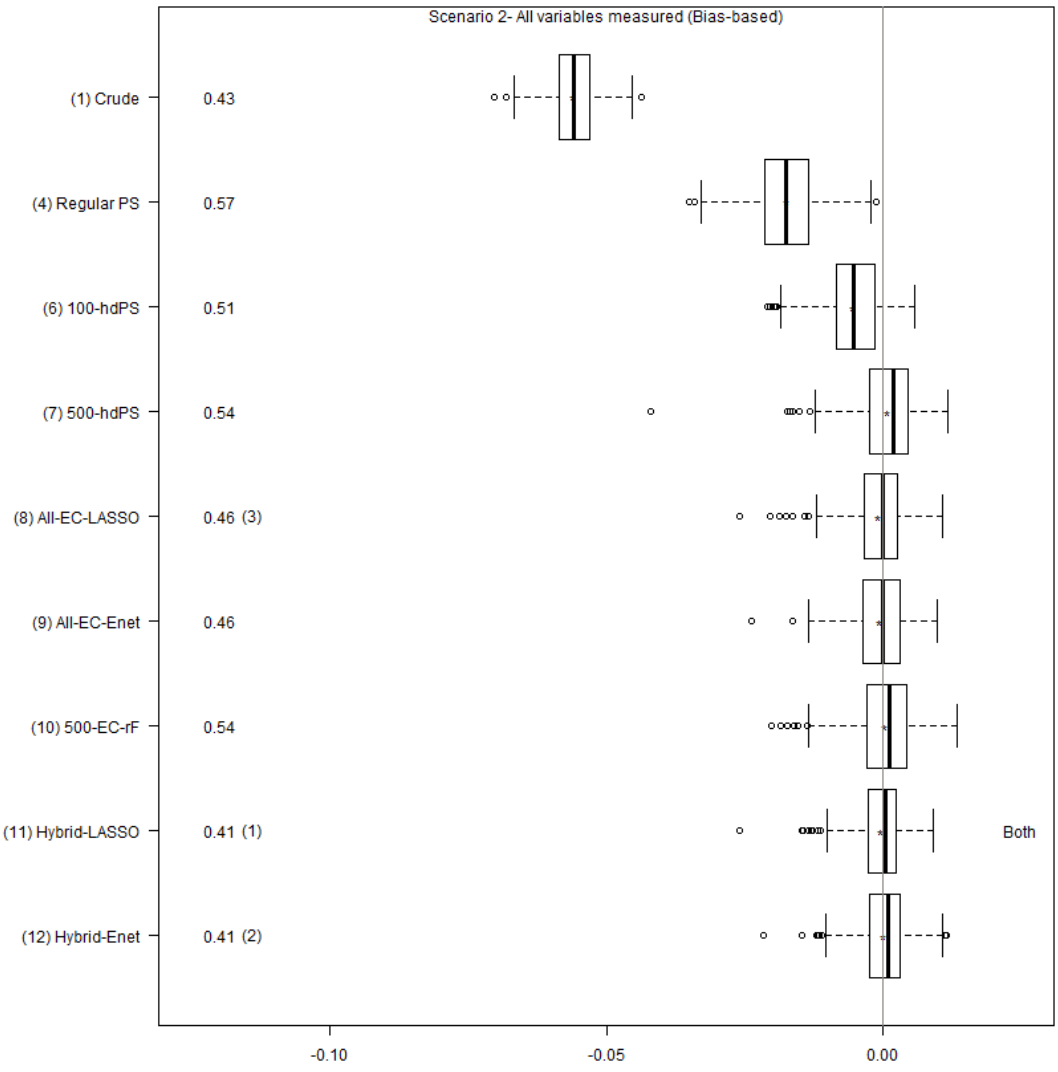


eFigure A.25: Plasmode Simulation Scenario 9-A

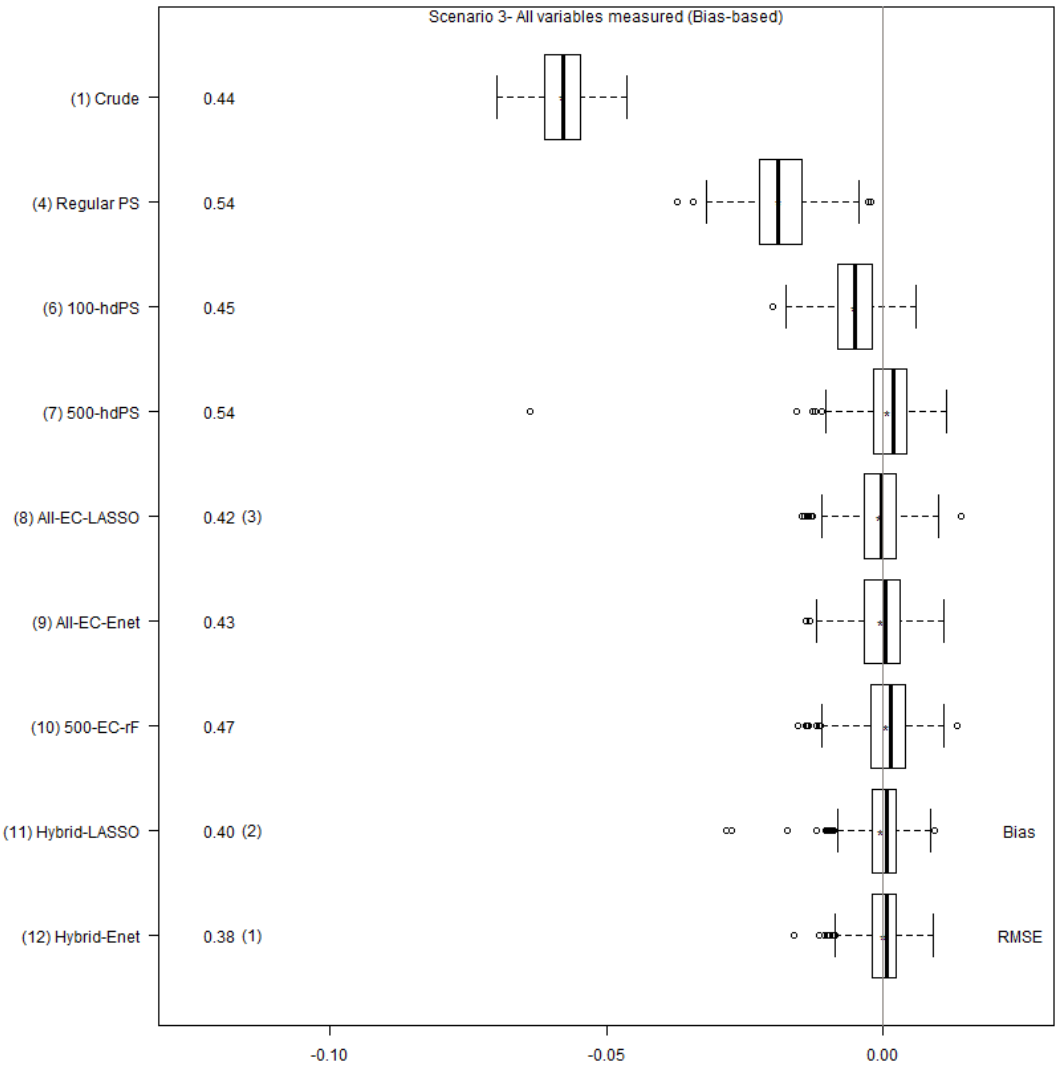
A.9.4 If all variables accounted (Bias-based analysis)



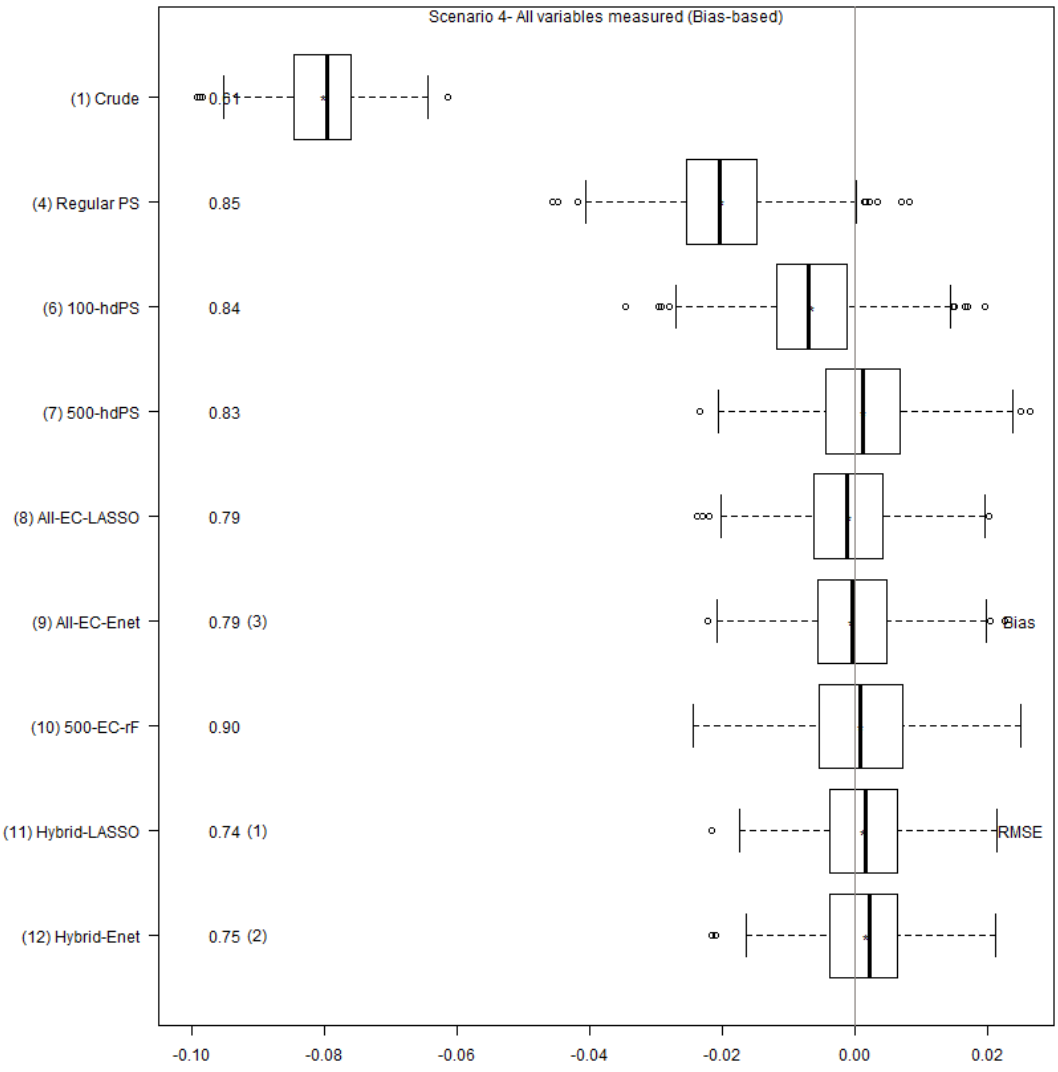
eFigure A.26: Plasmode Simulation Scenario 1-A



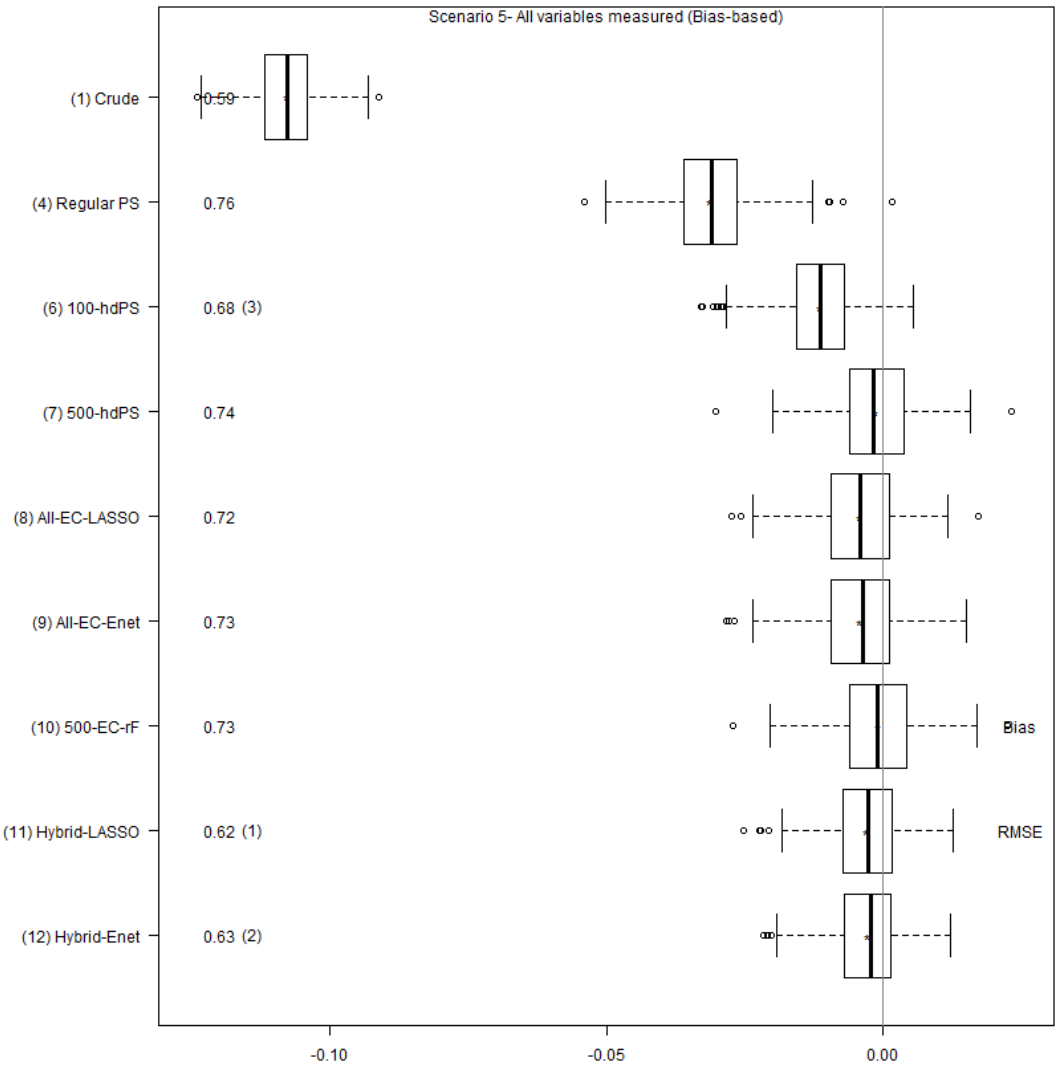
eFigure A.27: Plasmode Simulation Scenario 2-A



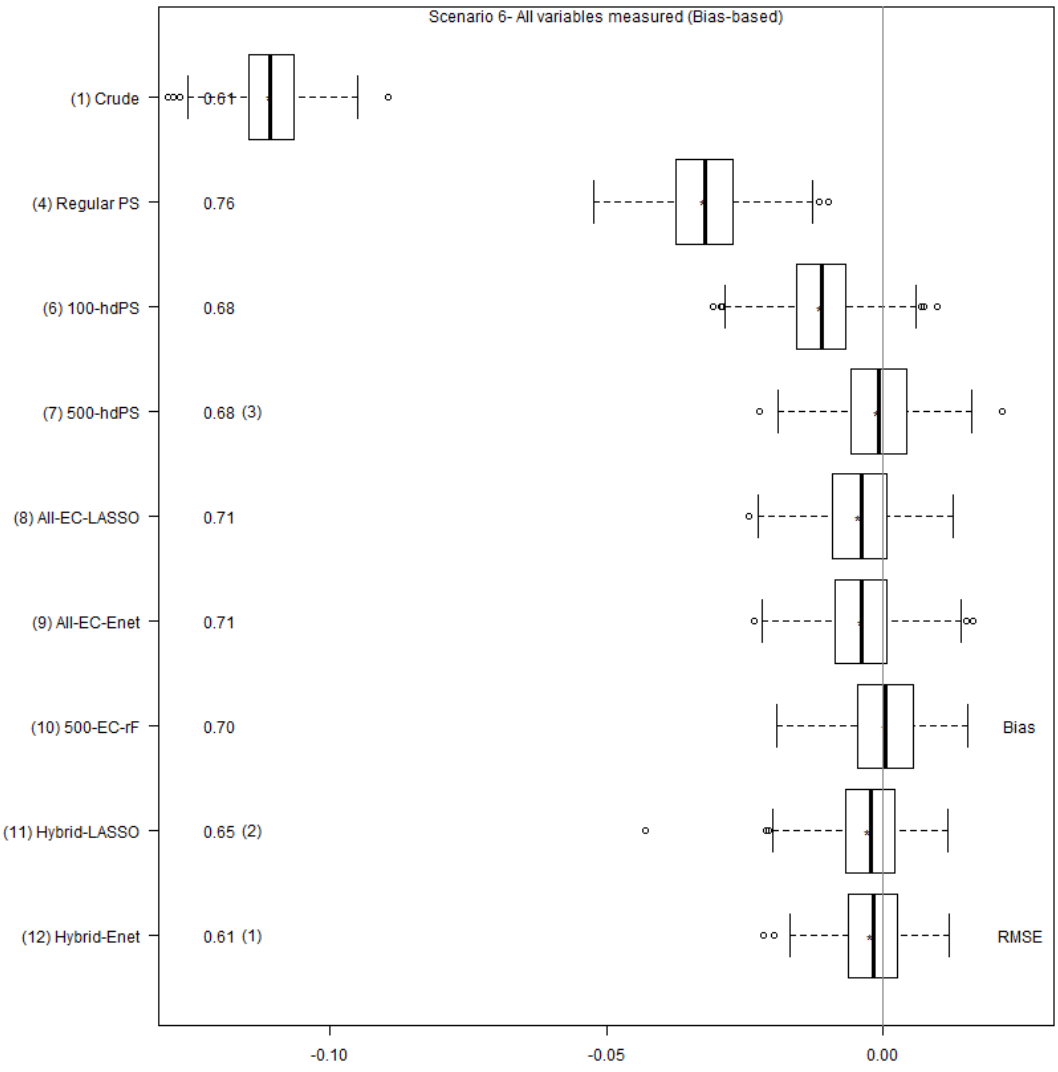
eFigure A.28: Plasmode Simulation Scenario 3-A



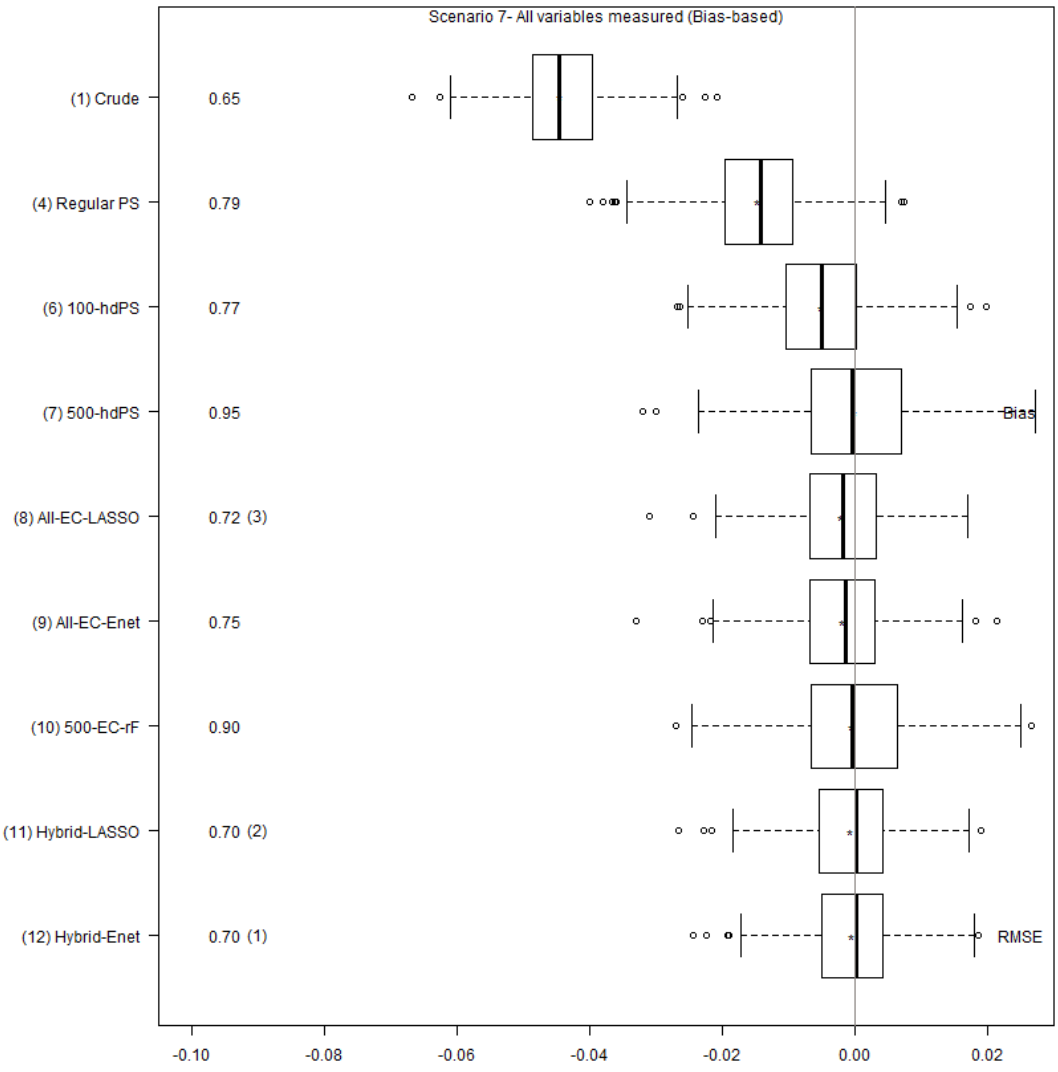
eFigure A.29: Plasmode Simulation Scenario 4-A



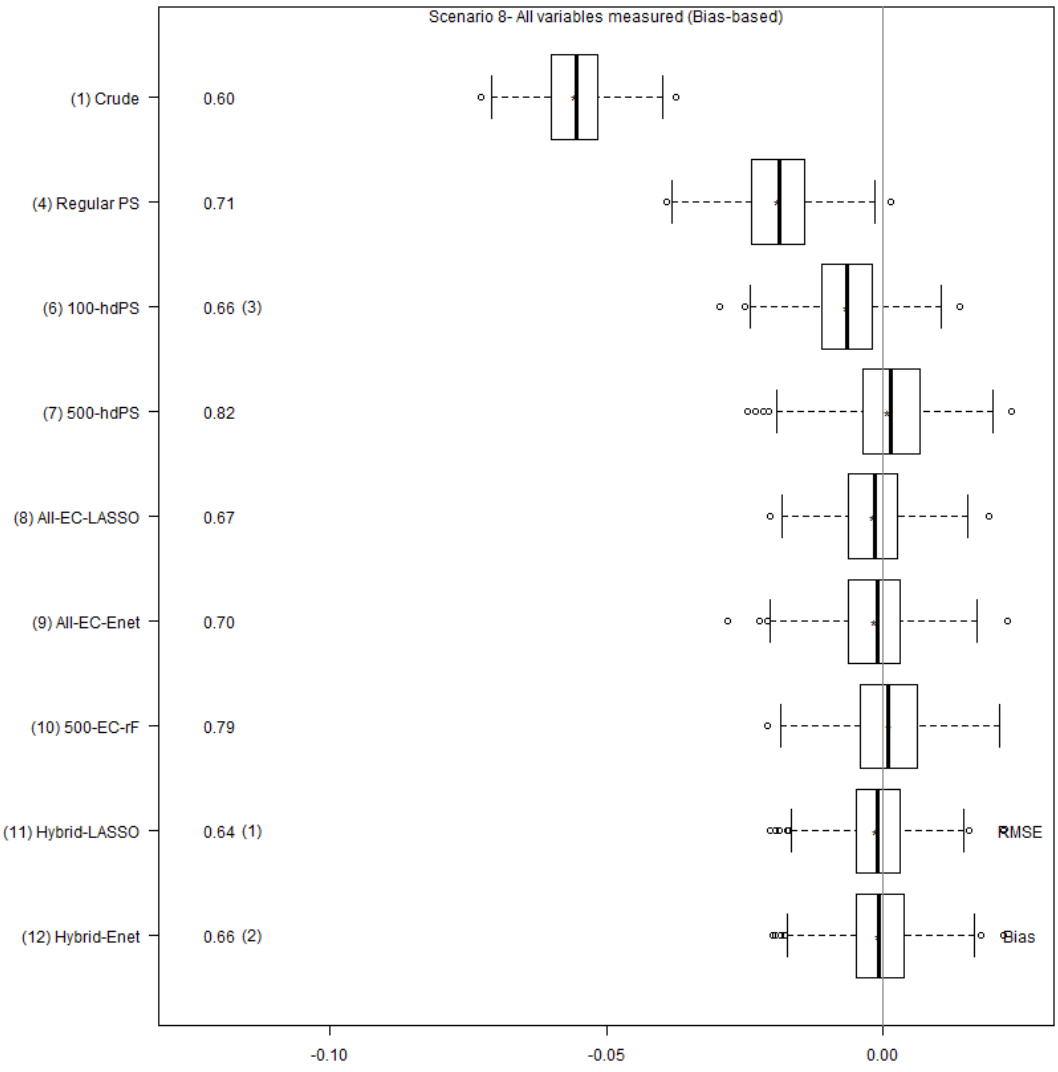
eFigure A.30: Plasmode Simulation Scenario 5-A



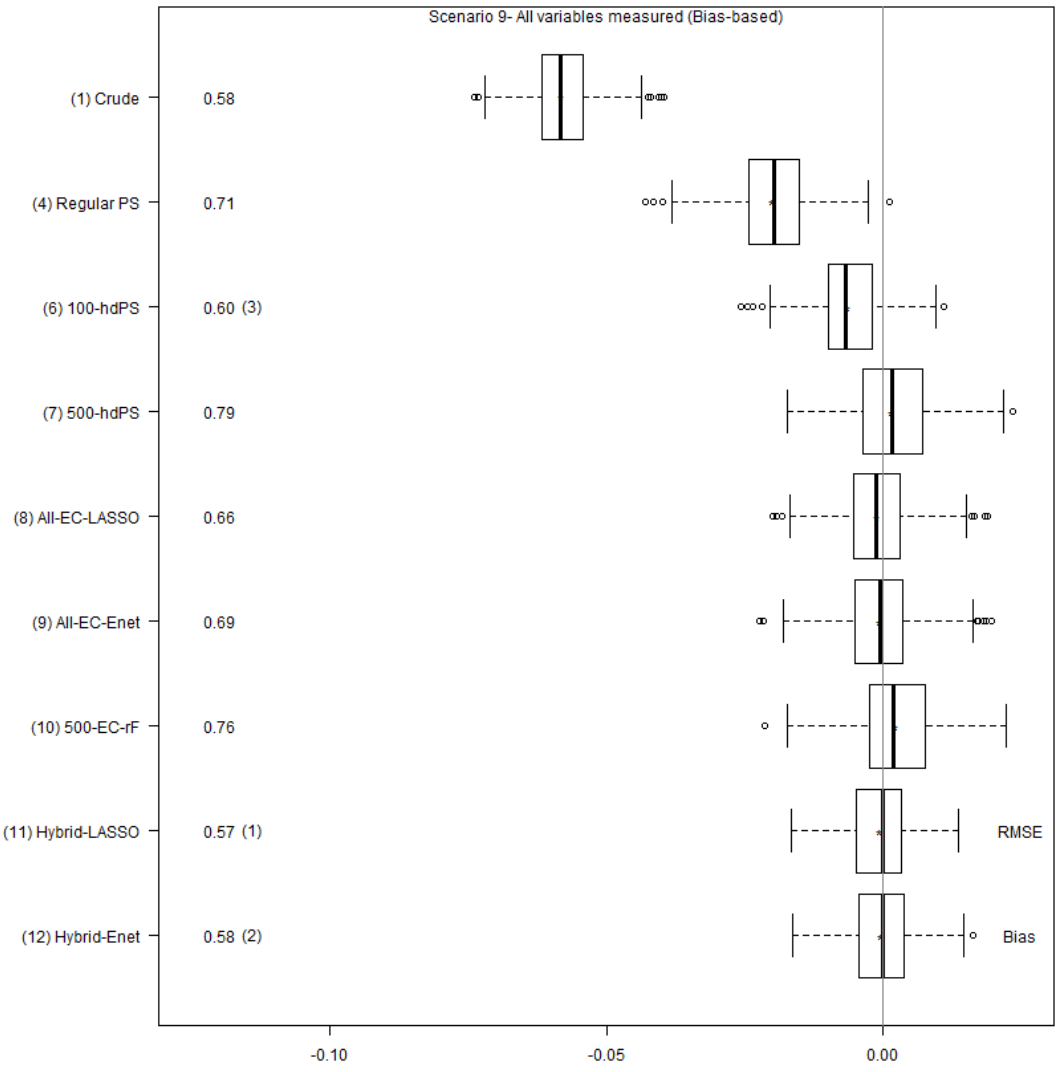
eFigure A.31: Plasmode Simulation Scenario 6-A



eFigure A.32: Plasmode Simulation Scenario 7-A

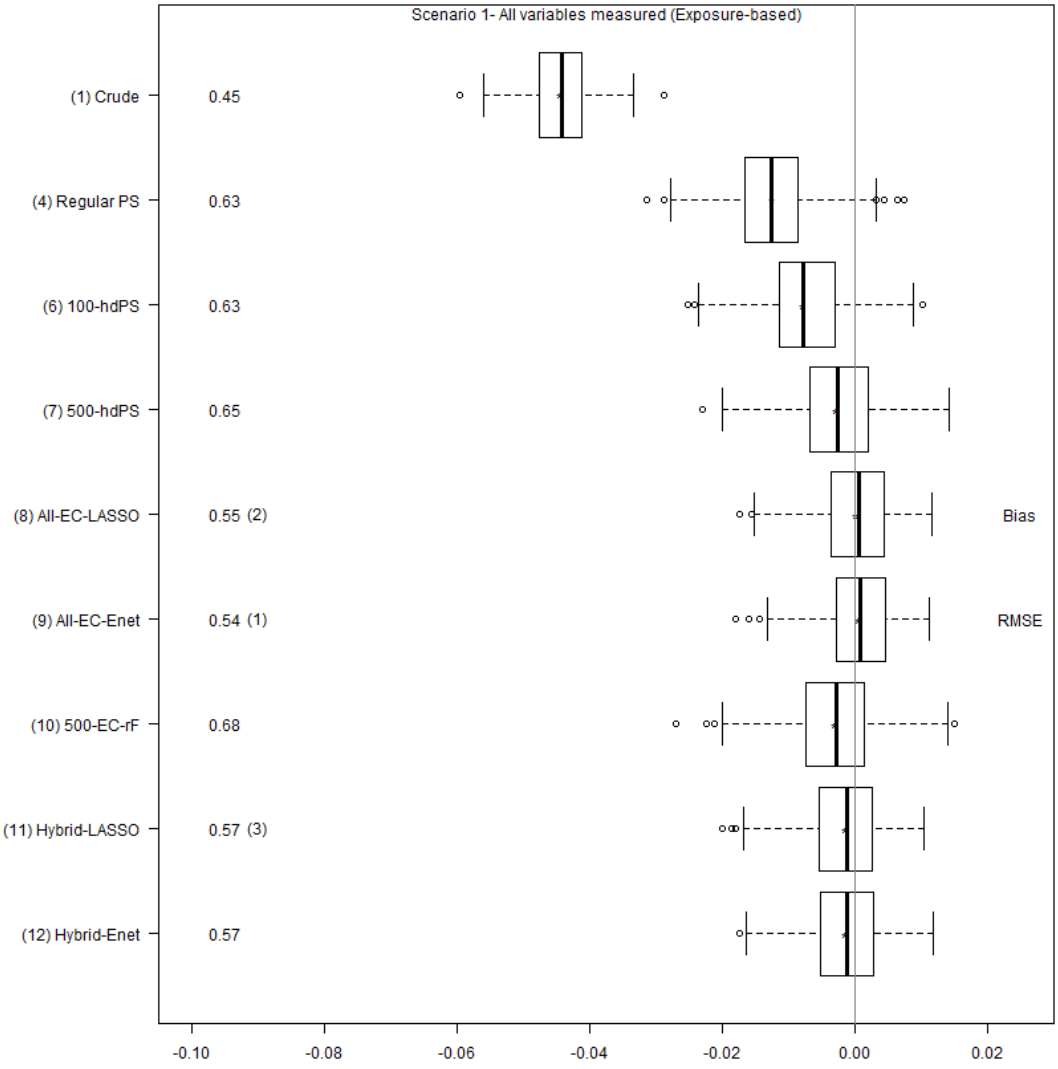


eFigure A.33: Plasmode Simulation Scenario 8-A

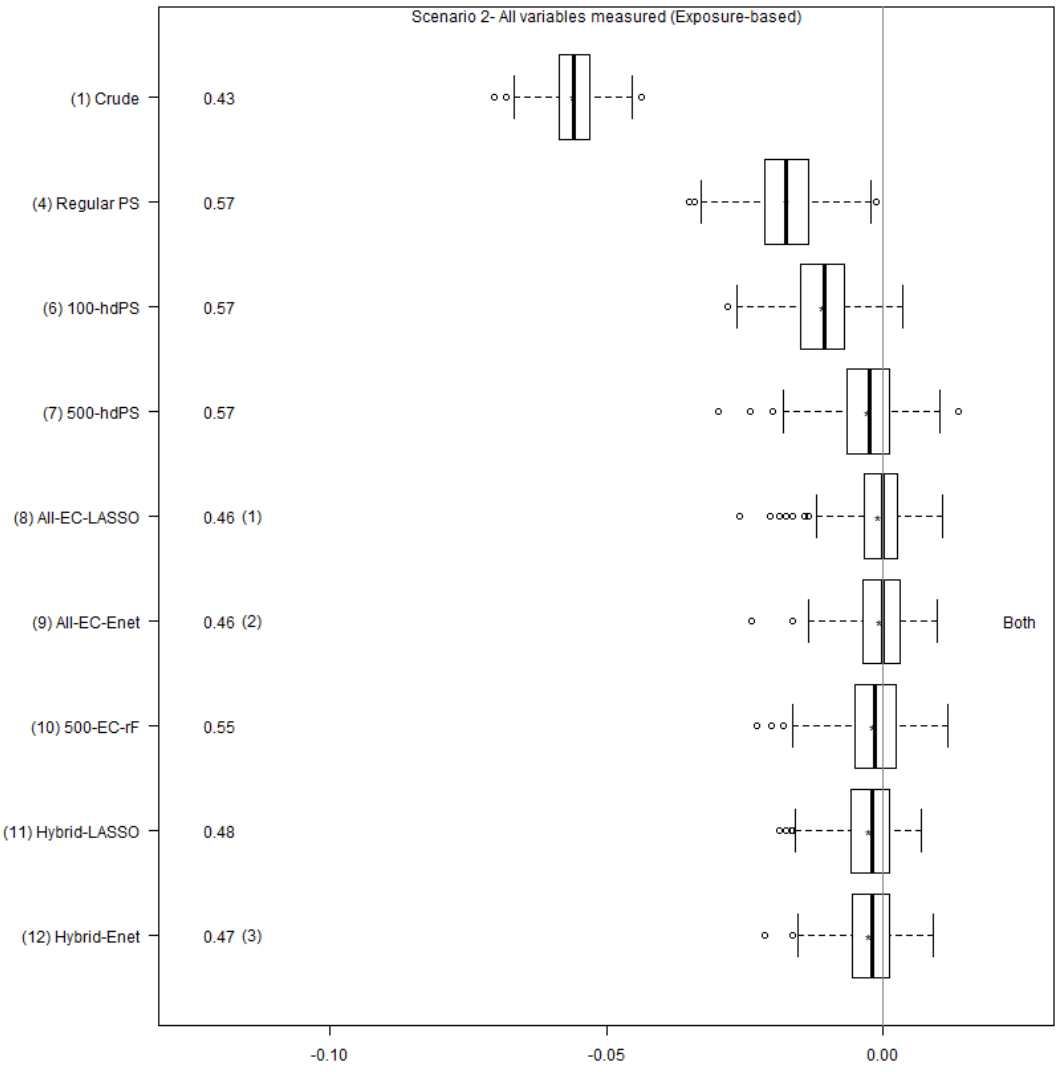


eFigure A.34: Plasmode Simulation Scenario 9-A

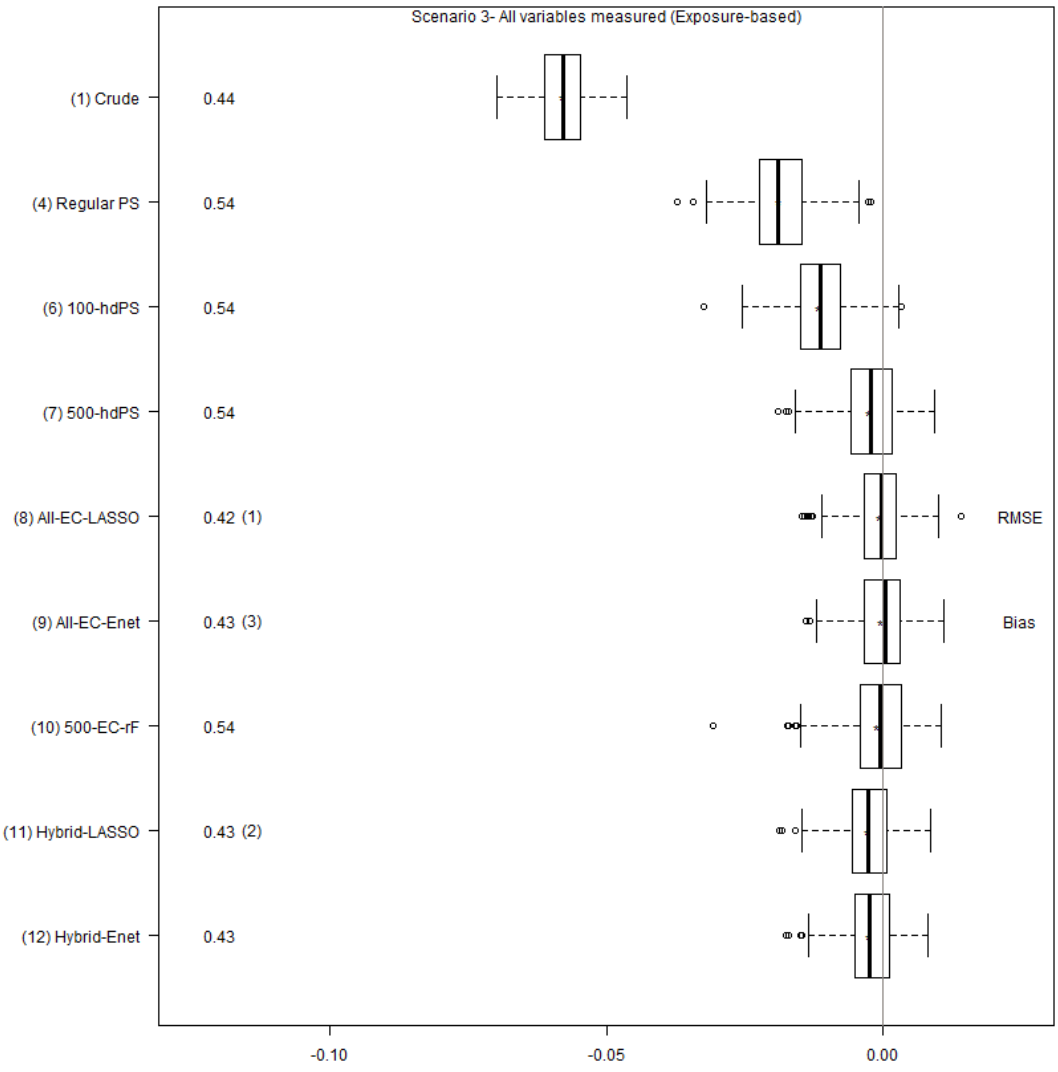
A.9.5 If all variables accounted (Exposure-based analysis)



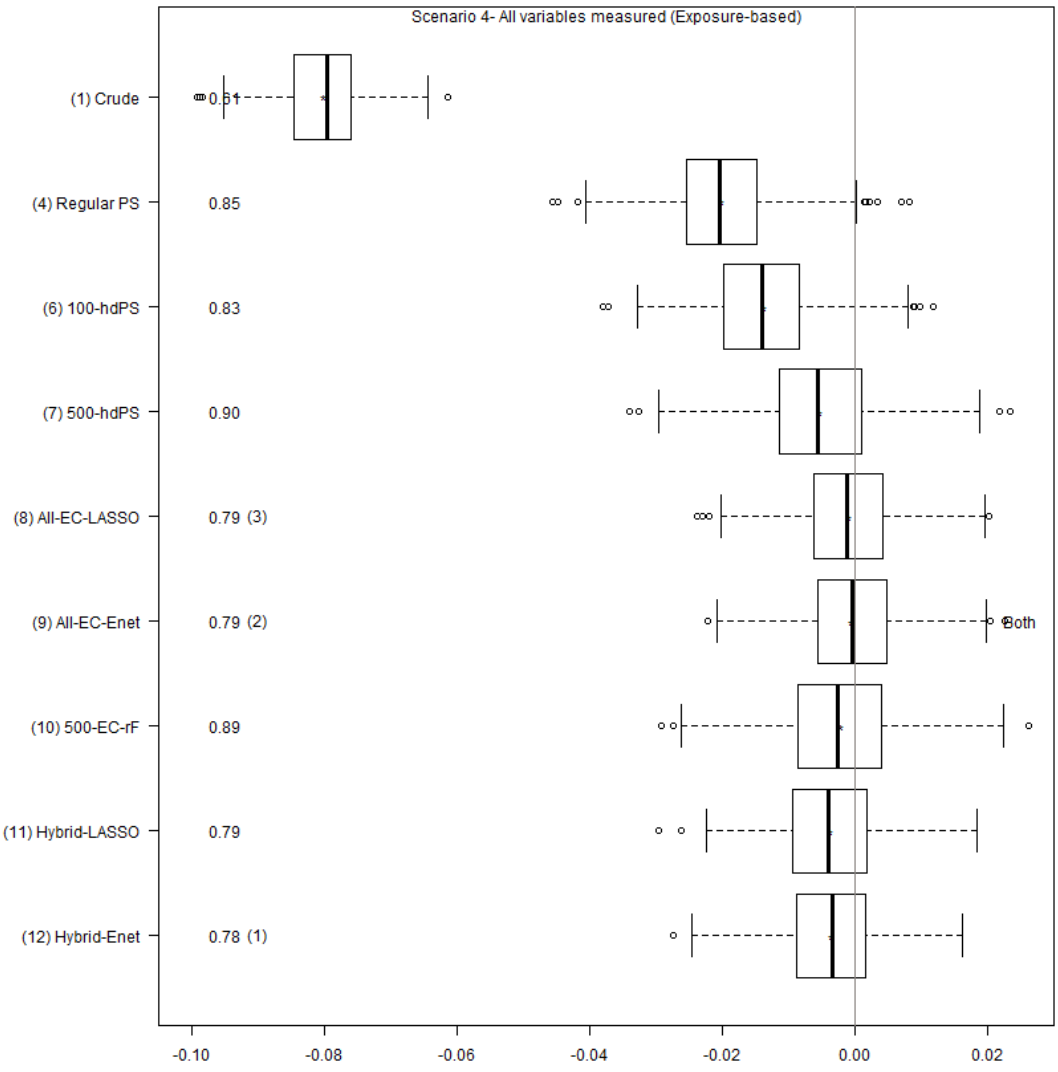
eFigure A.35: Plasmode Simulation Scenario 1-A



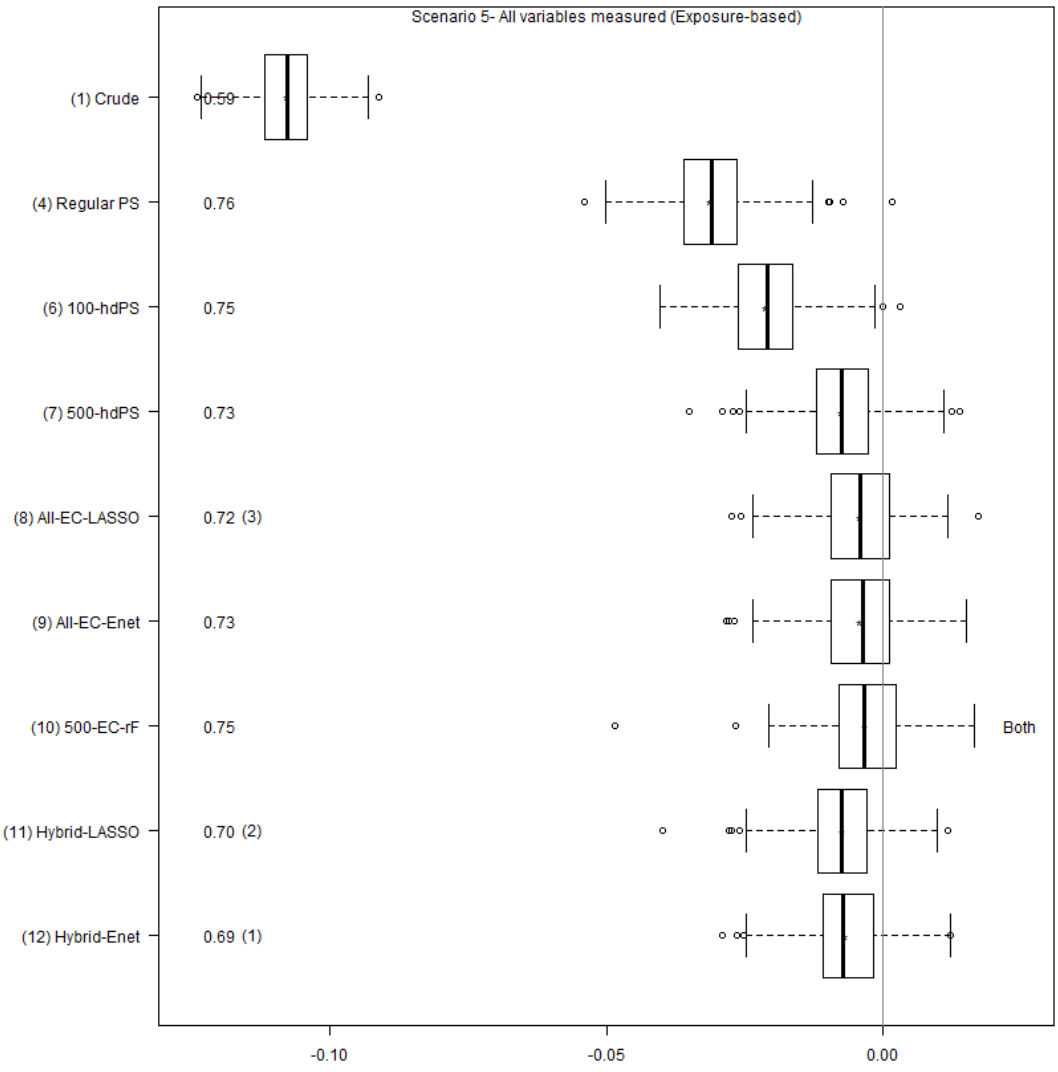
eFigure A.36: Plasmode Simulation Scenario 2-A



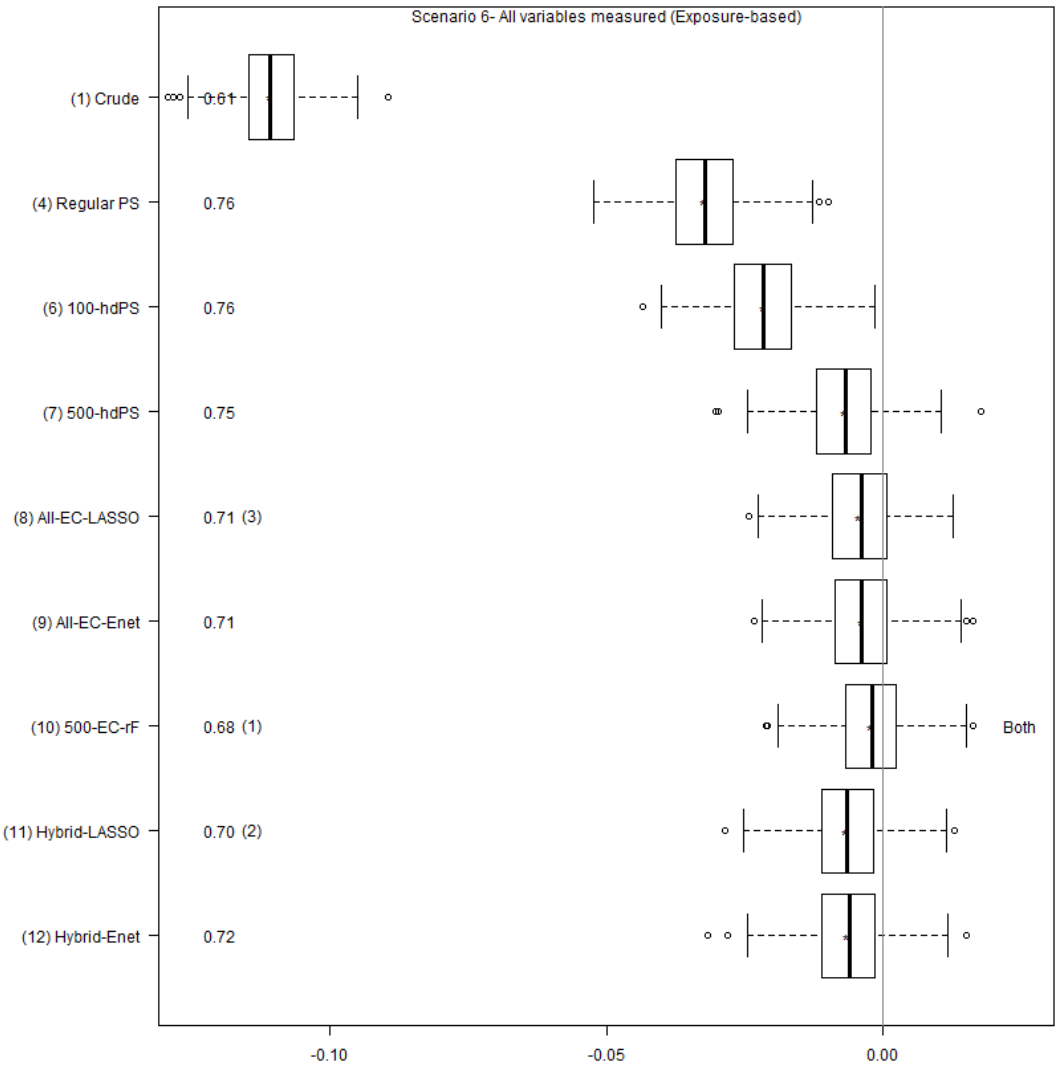
eFigure A.37: Plasmode Simulation Scenario 3-A



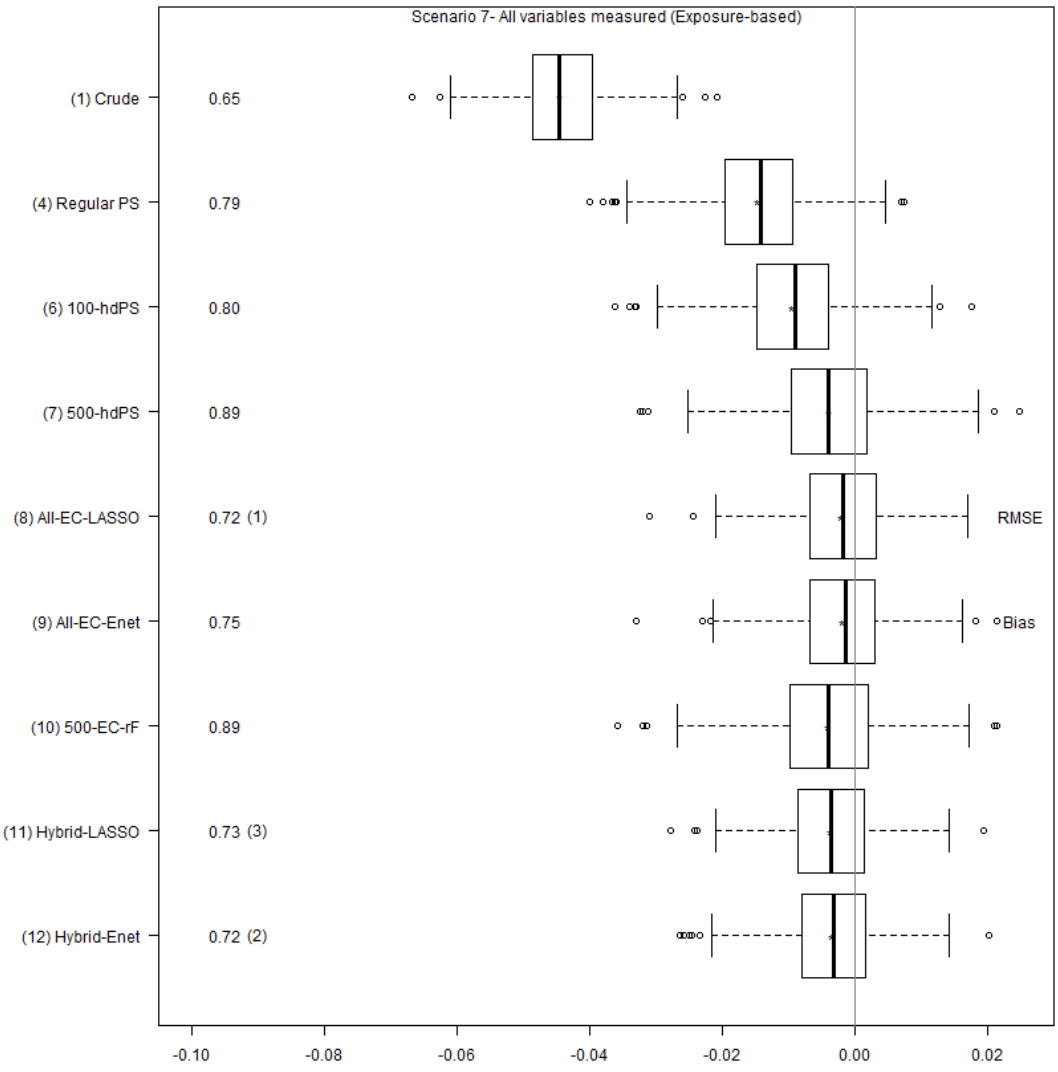
eFigure A.38: Plasmode Simulation Scenario 4-A



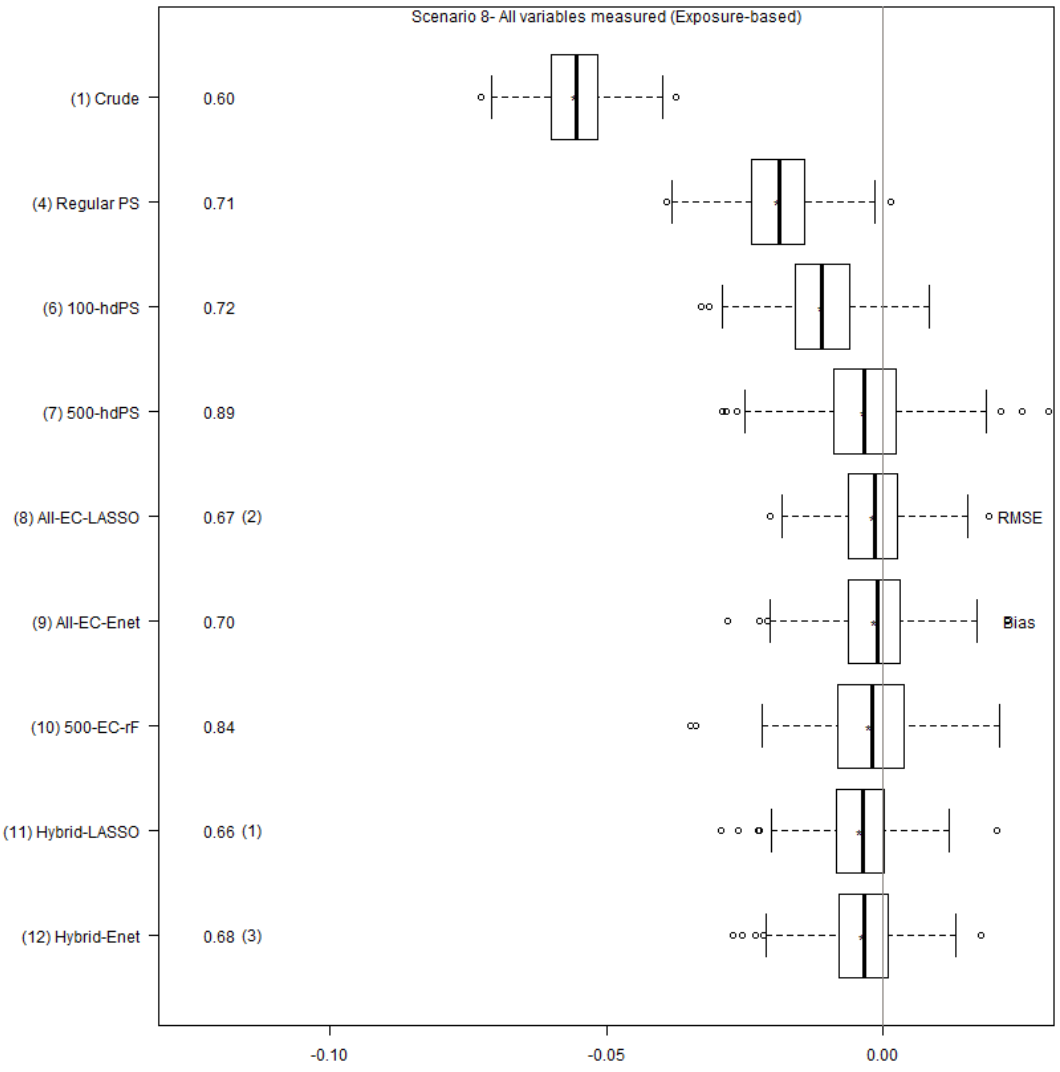
eFigure A.39: Plasmode Simulation Scenario 5-A



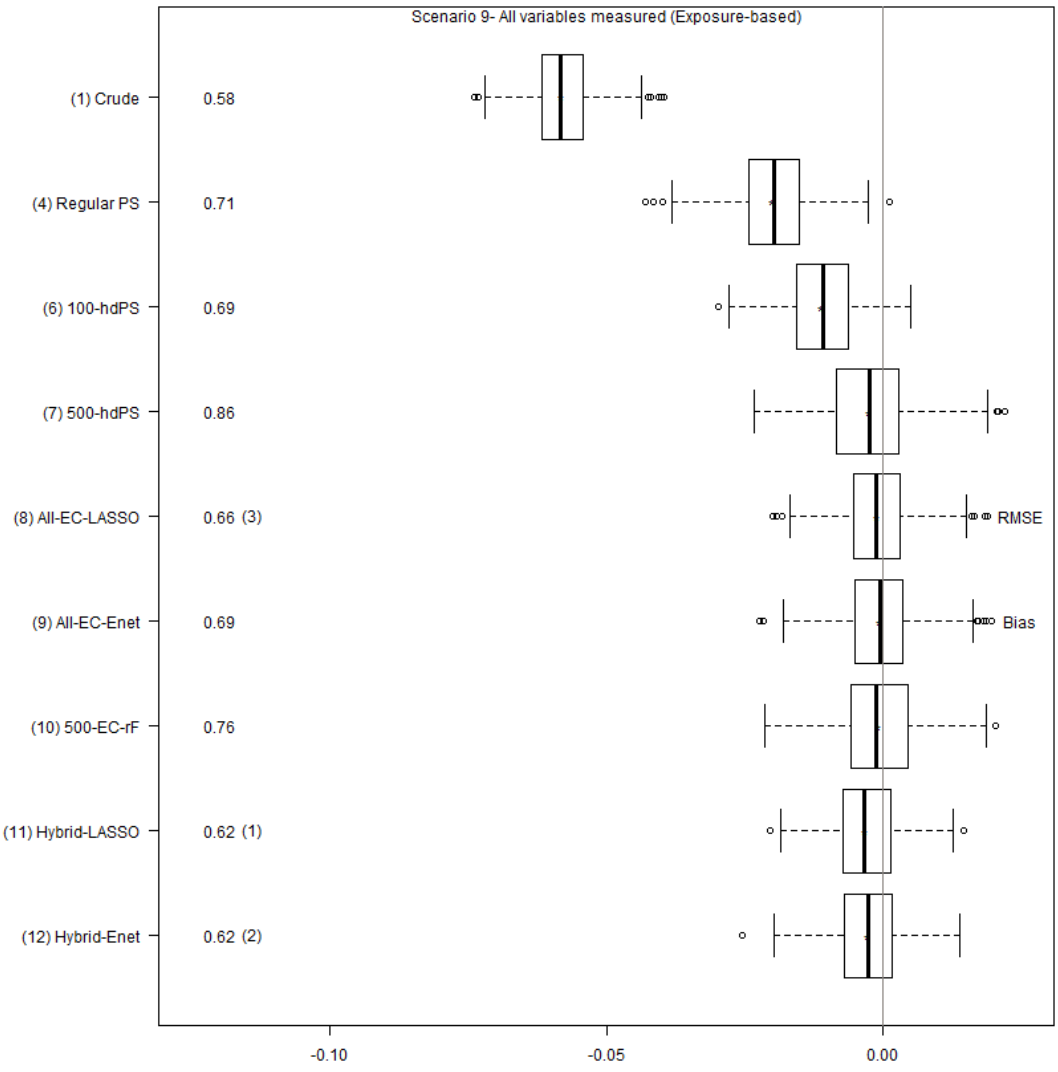
eFigure A.40: Plasmode Simulation Scenario 6-A



eFigure A.41: Plasmode Simulation Scenario 7-A



eFigure A.42: Plasmode Simulation Scenario 8-A

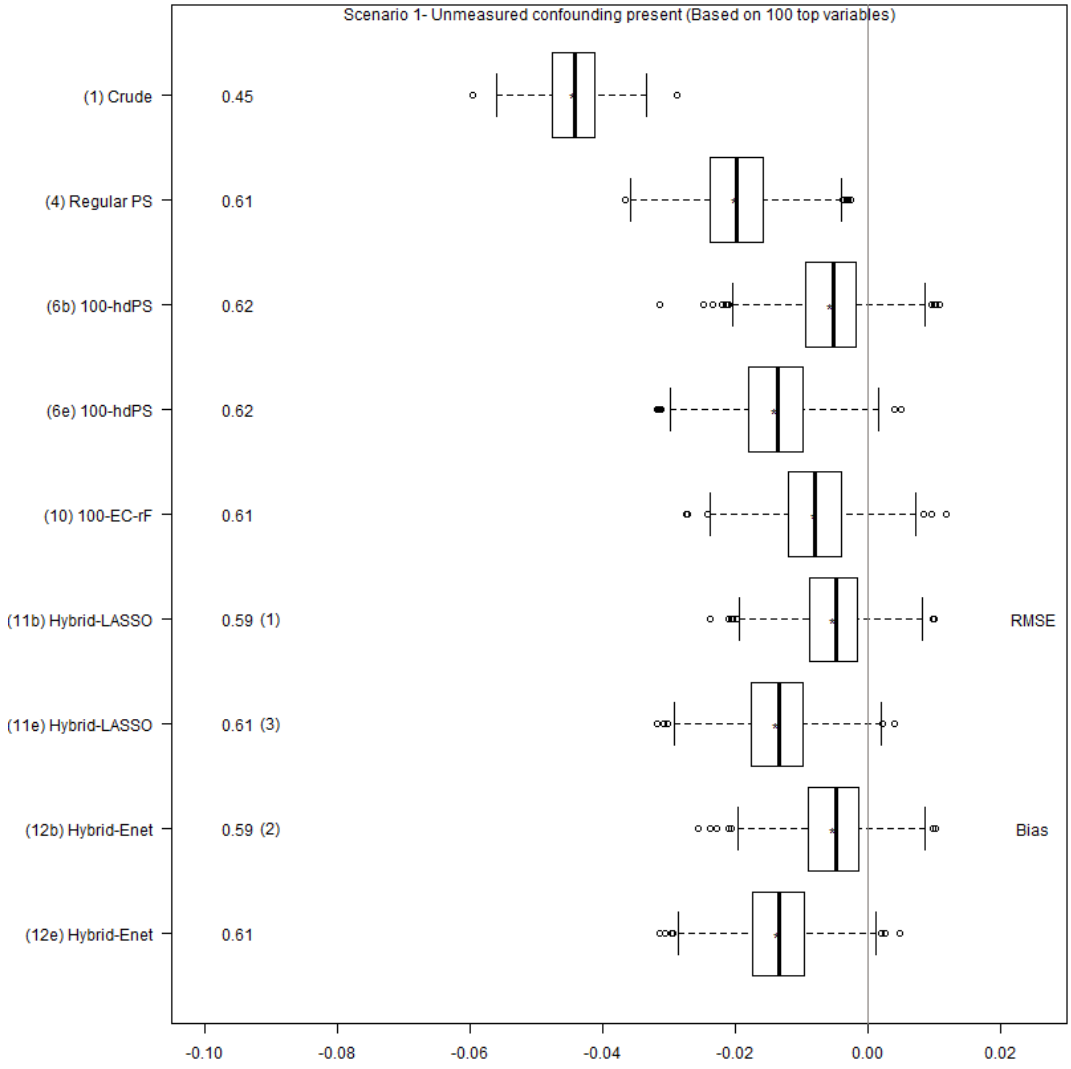


eFigure A.43: Plasmode Simulation Scenario 9-A

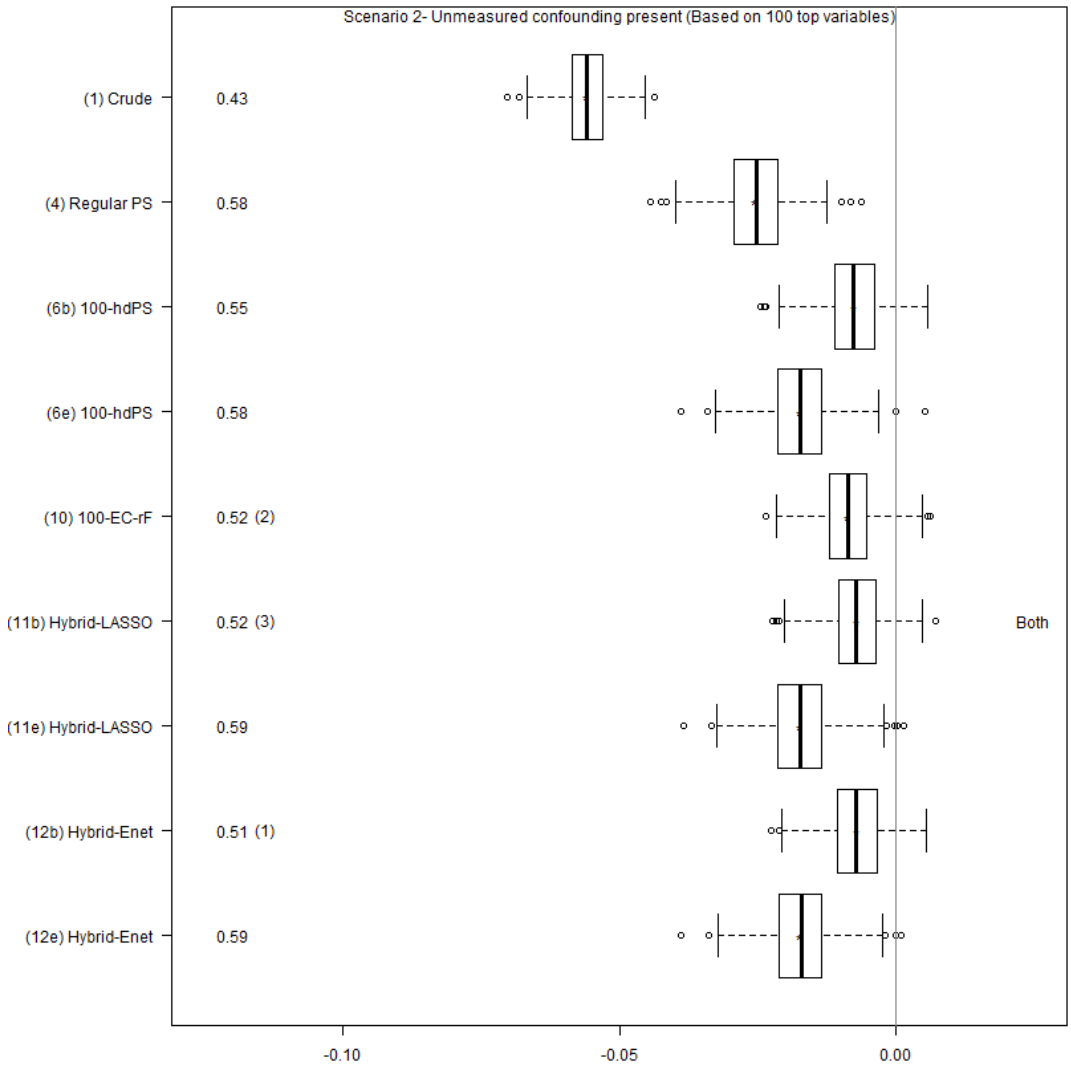
A.9.6 Considering fewer variables in the analysis

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.

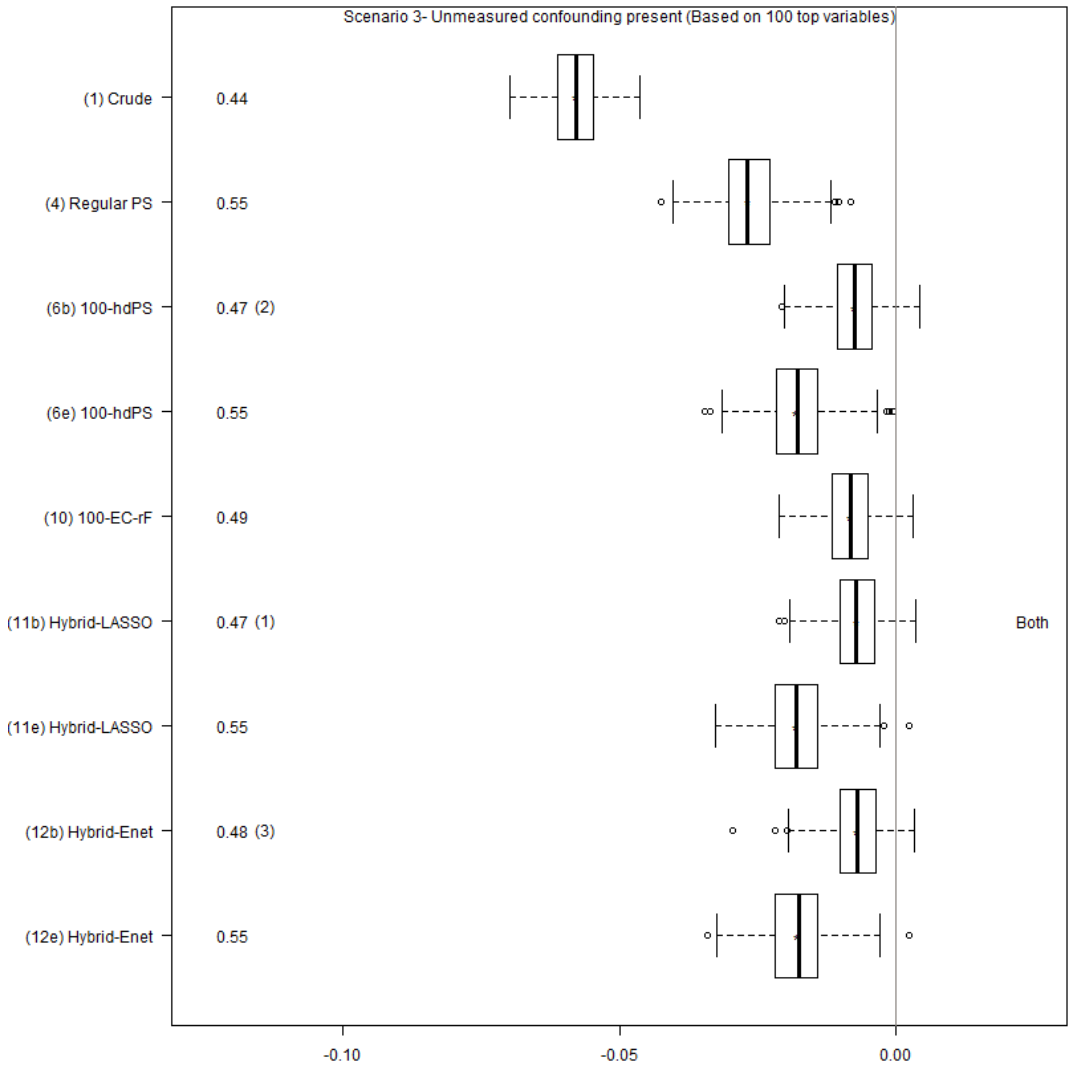
A.9.7 If unmeasured confounding present (Based on top 100 selected variables)



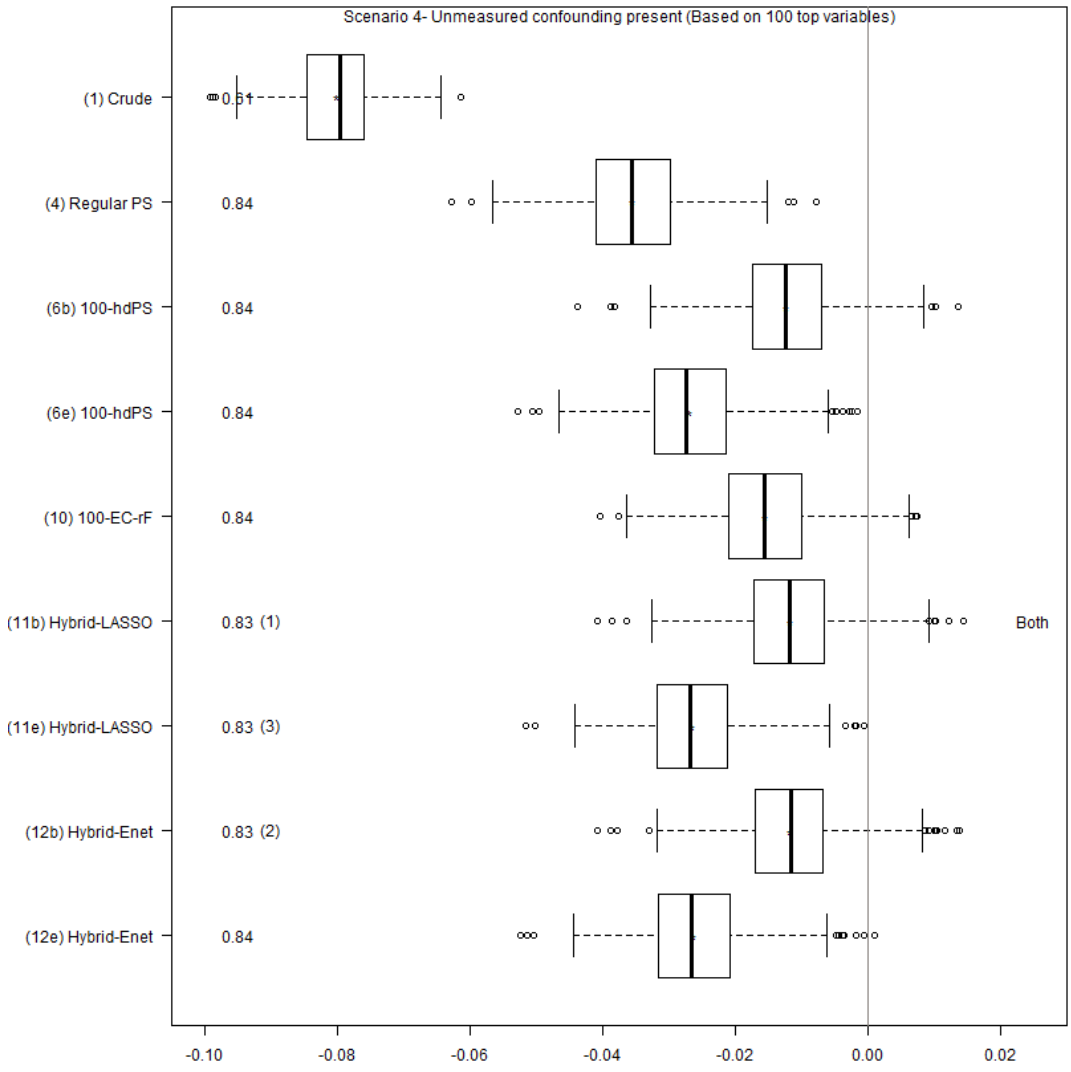
eFigure A.44: Plasmode Simulation Scenario 1-A



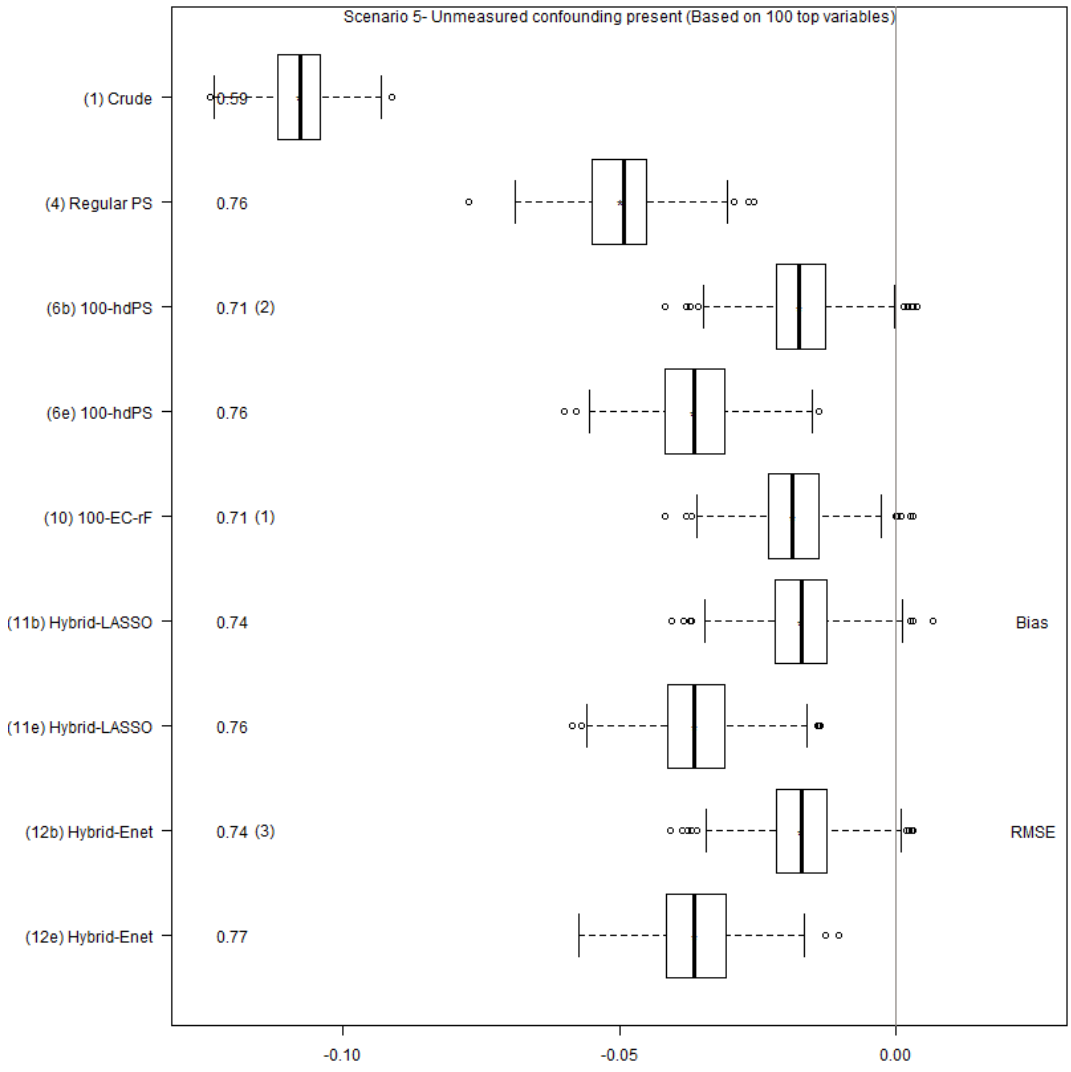
eFigure A.45: Plasmode Simulation Scenario 2-A



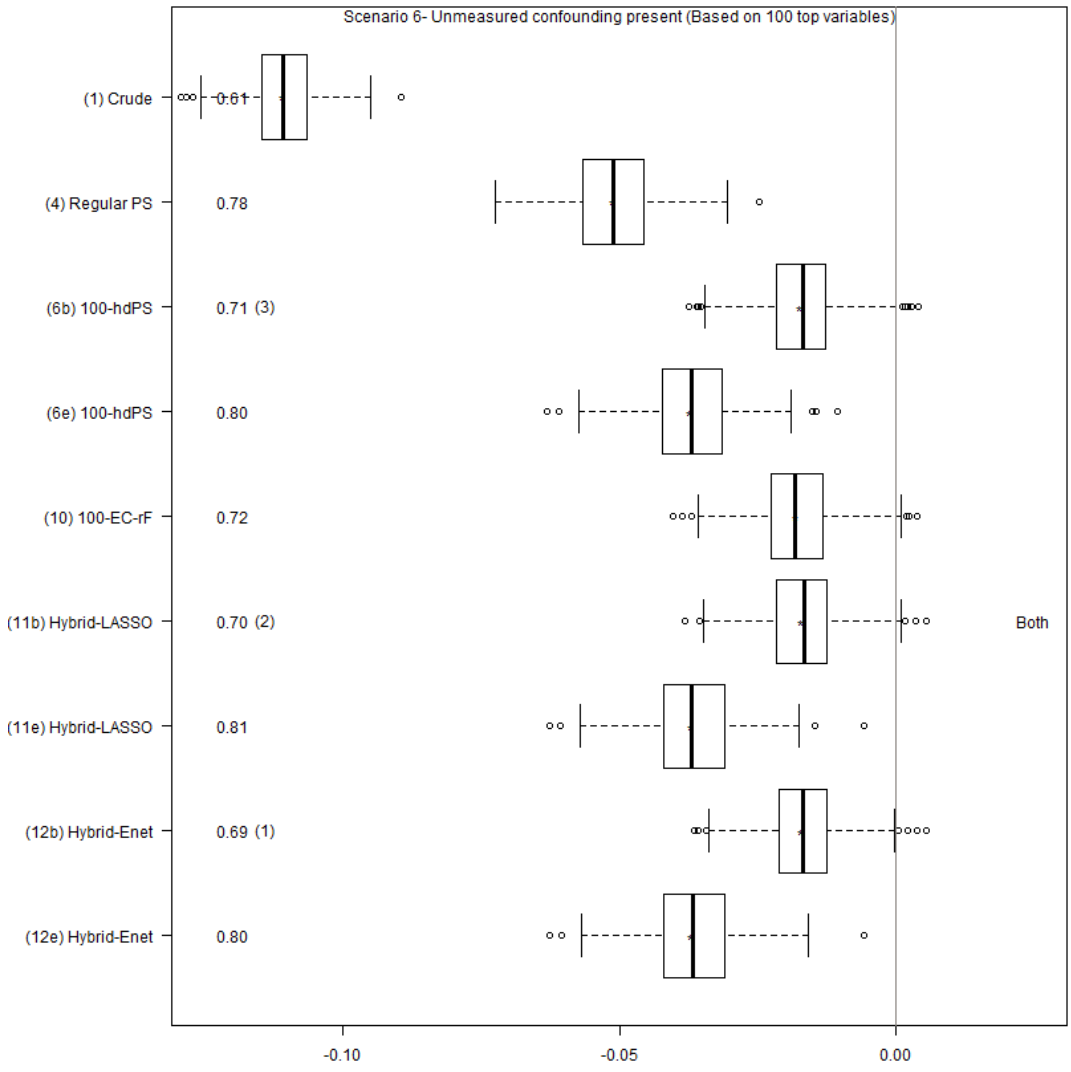
eFigure A.46: Plasmode Simulation Scenario 3-A



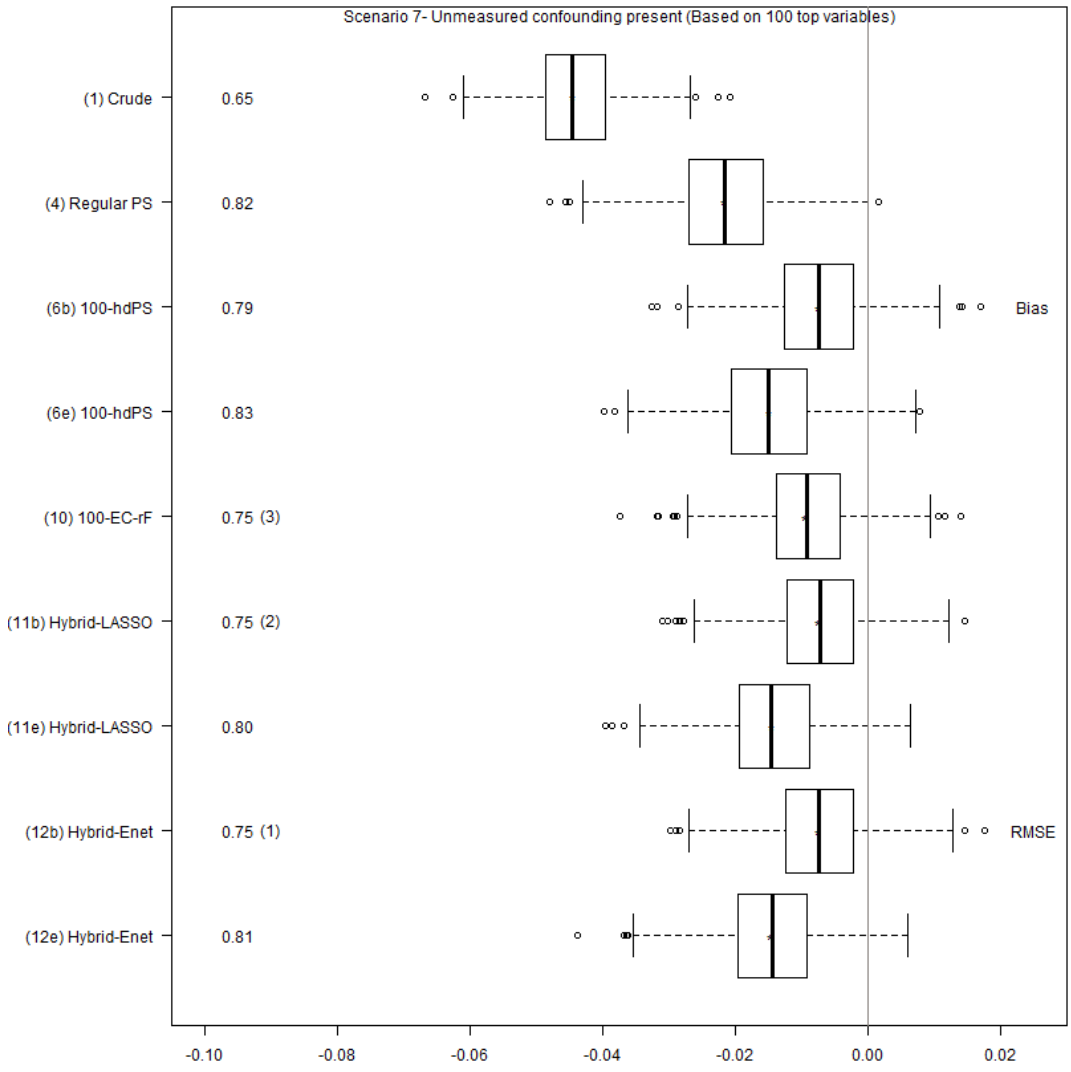
eFigure A.47: Plasmode Simulation Scenario 4-A



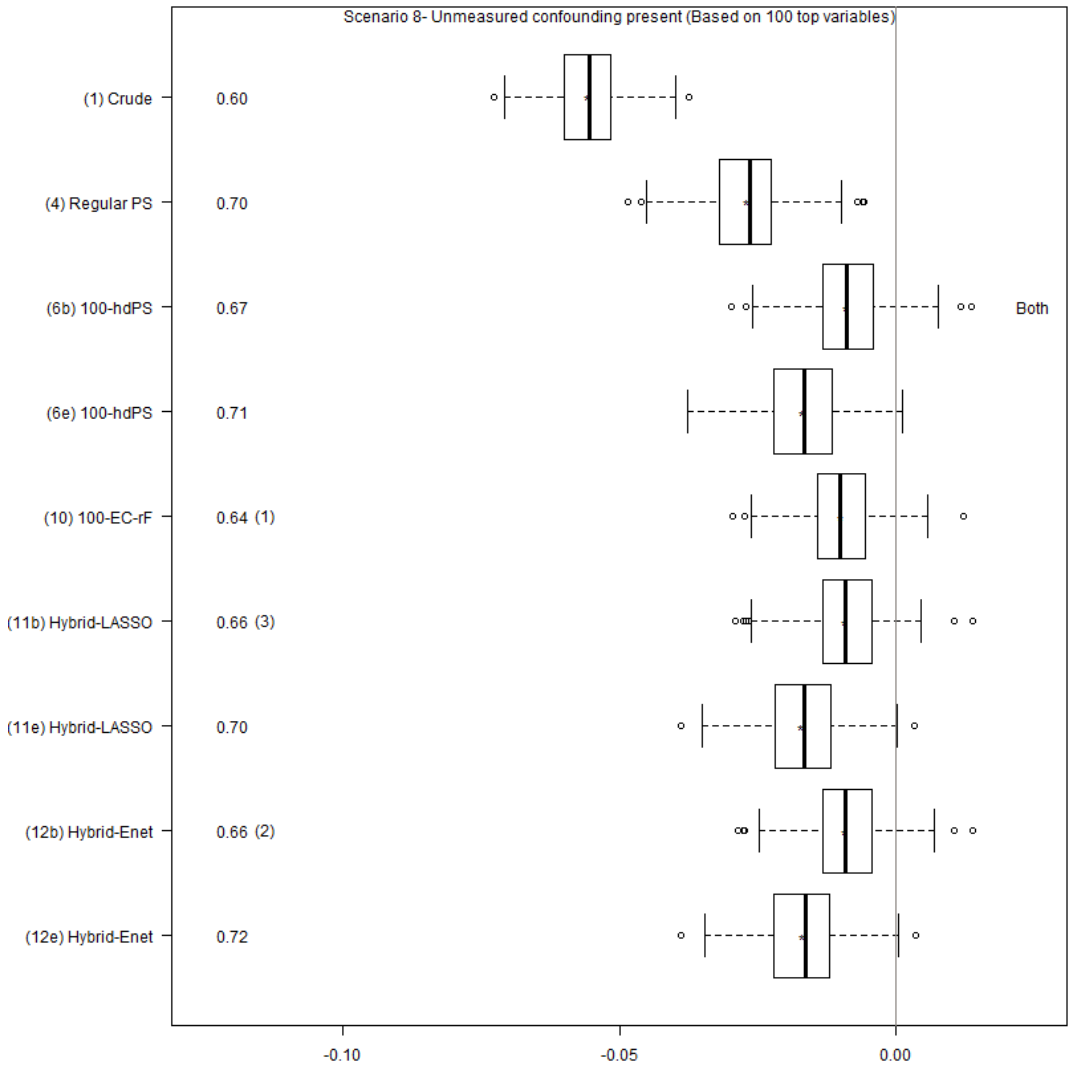
eFigure A.48: Plasmode Simulation Scenario 5-A



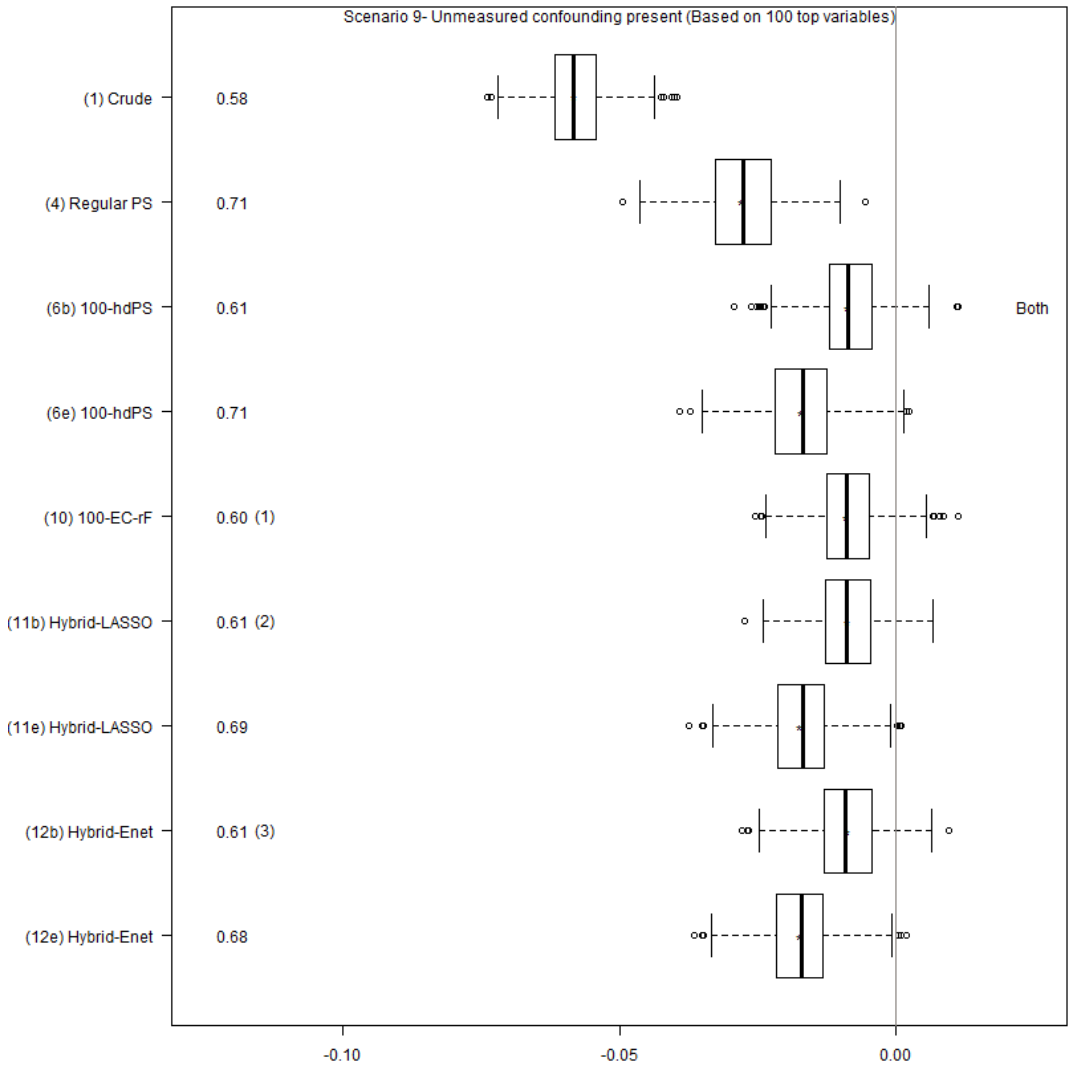
eFigure A.49: Plasmode Simulation Scenario 6-A



eFigure A.50: Plasmode Simulation Scenario 7-A

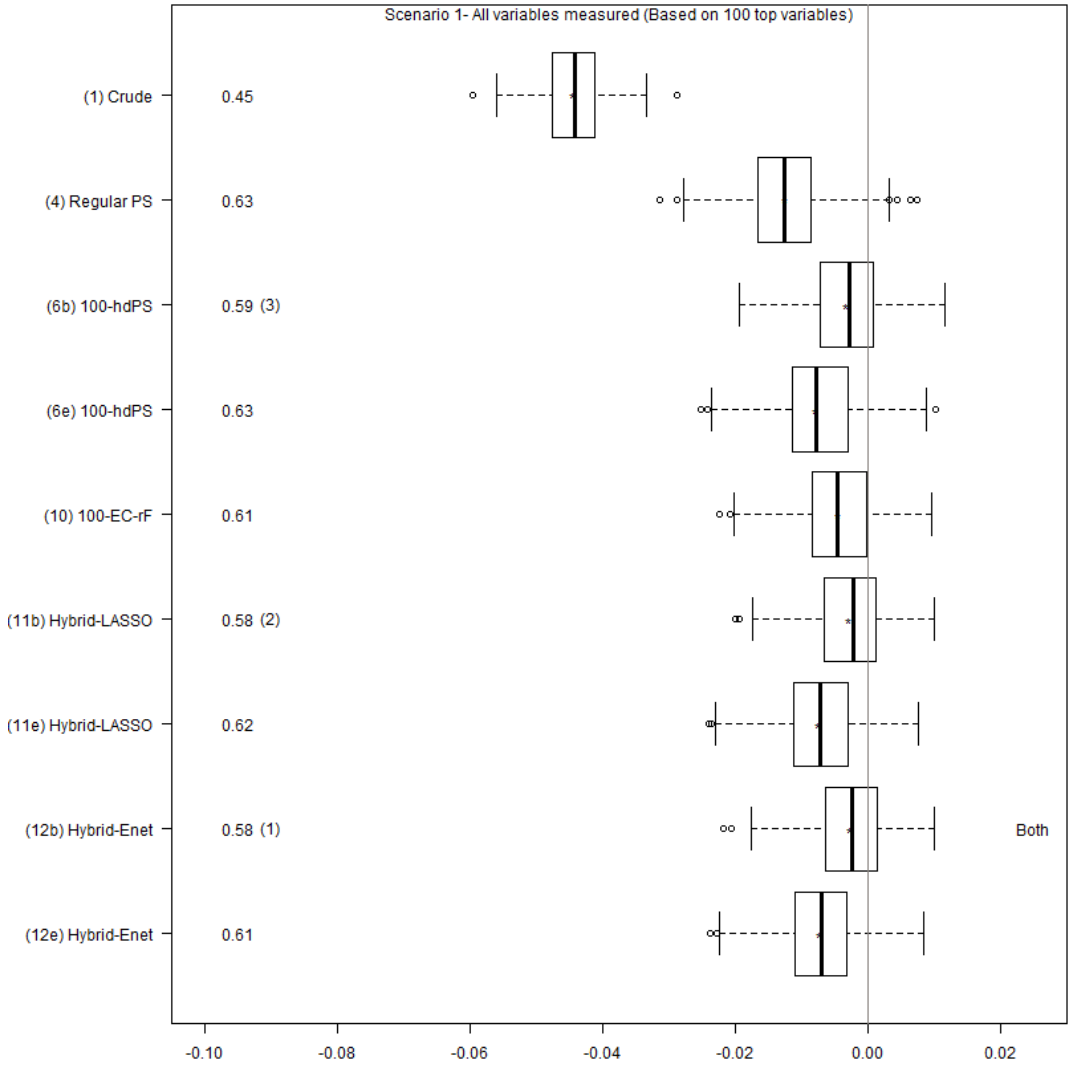


eFigure A.51: Plasmode Simulation Scenario 8-A

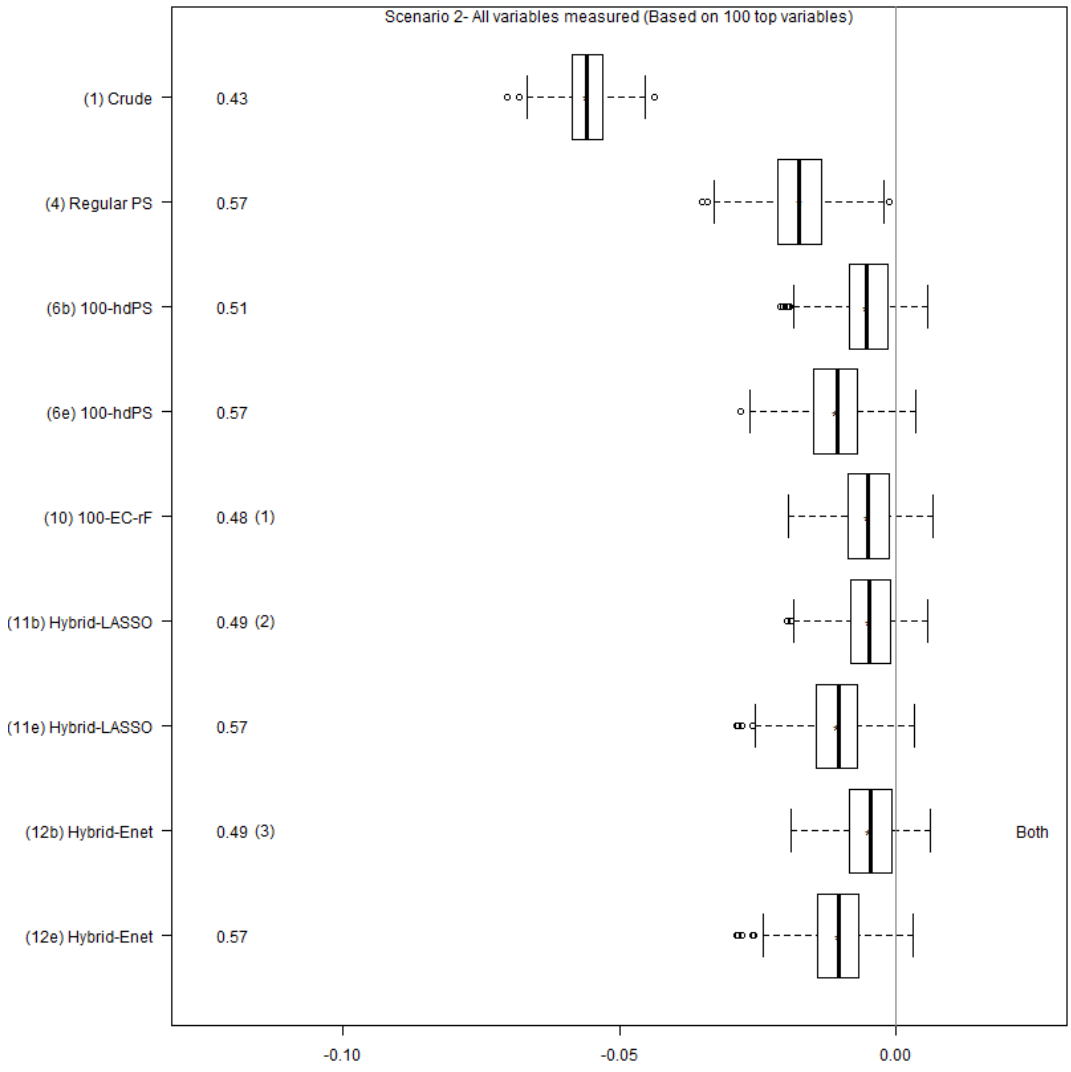


eFigure A.52: Plasmode Simulation Scenario 9-A

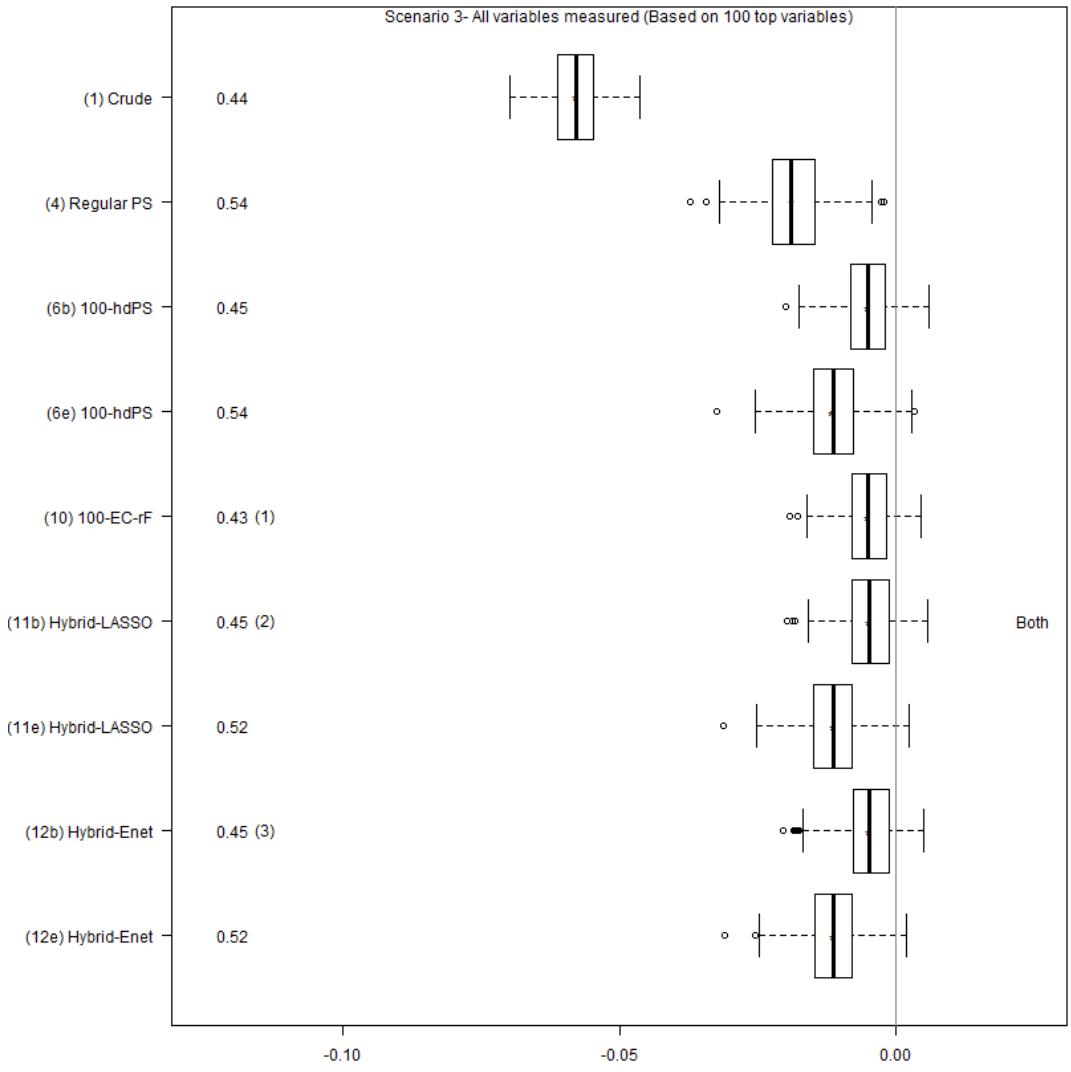
A.9.8 If all variables accounted (Based on top 100 selected variables)



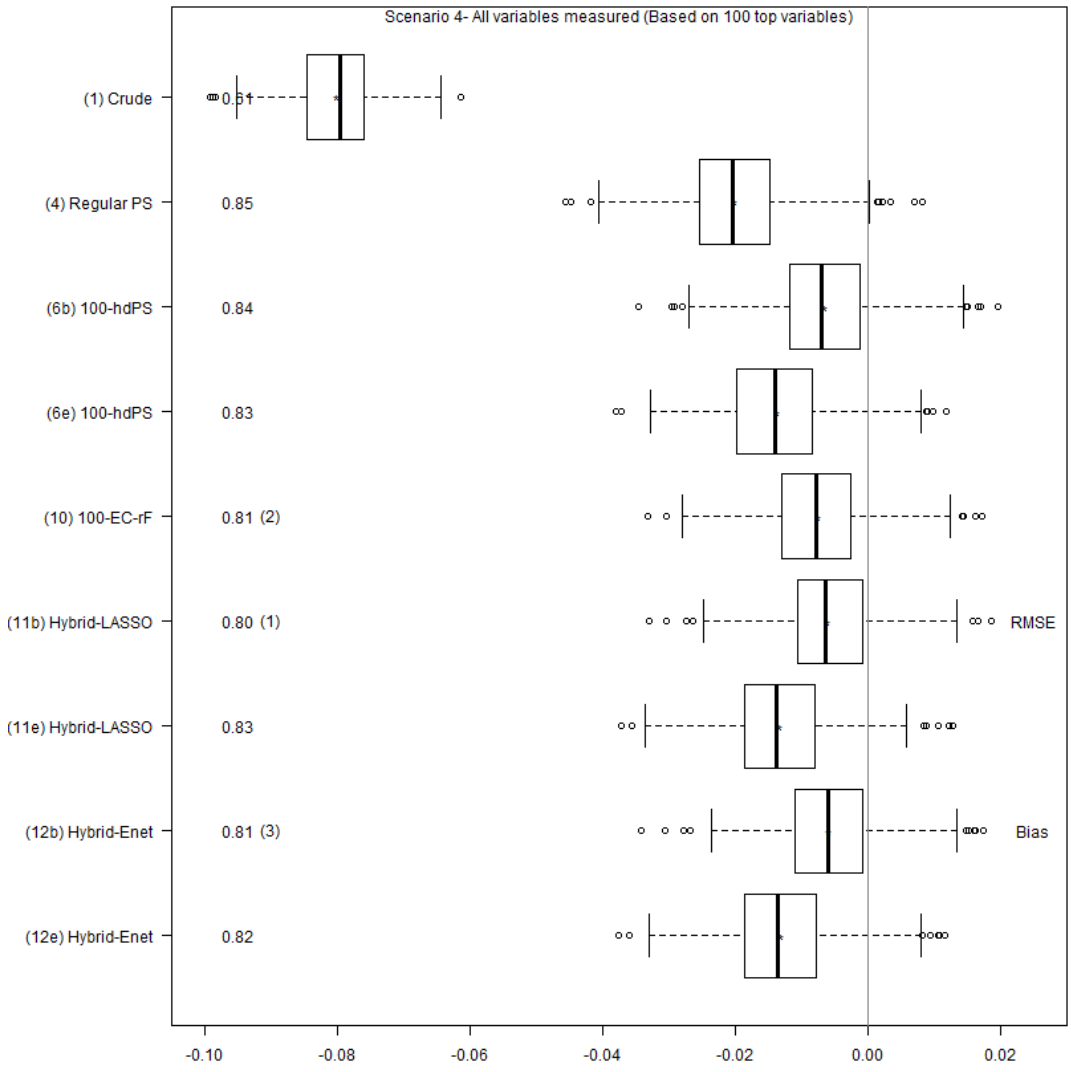
eFigure A.53: Plasmode Simulation Scenario 1-A



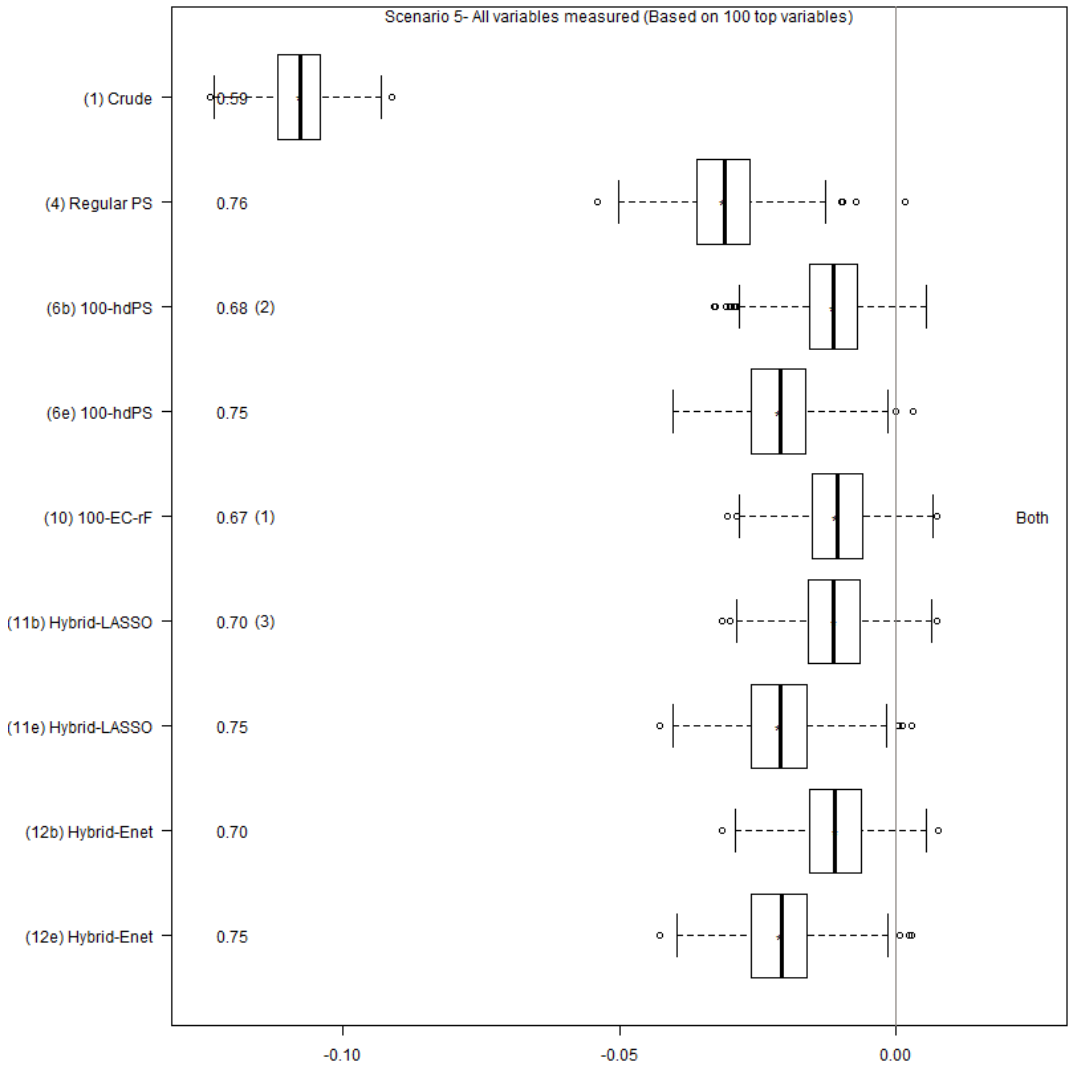
eFigure A.54: Plasmode Simulation Scenario 2-A



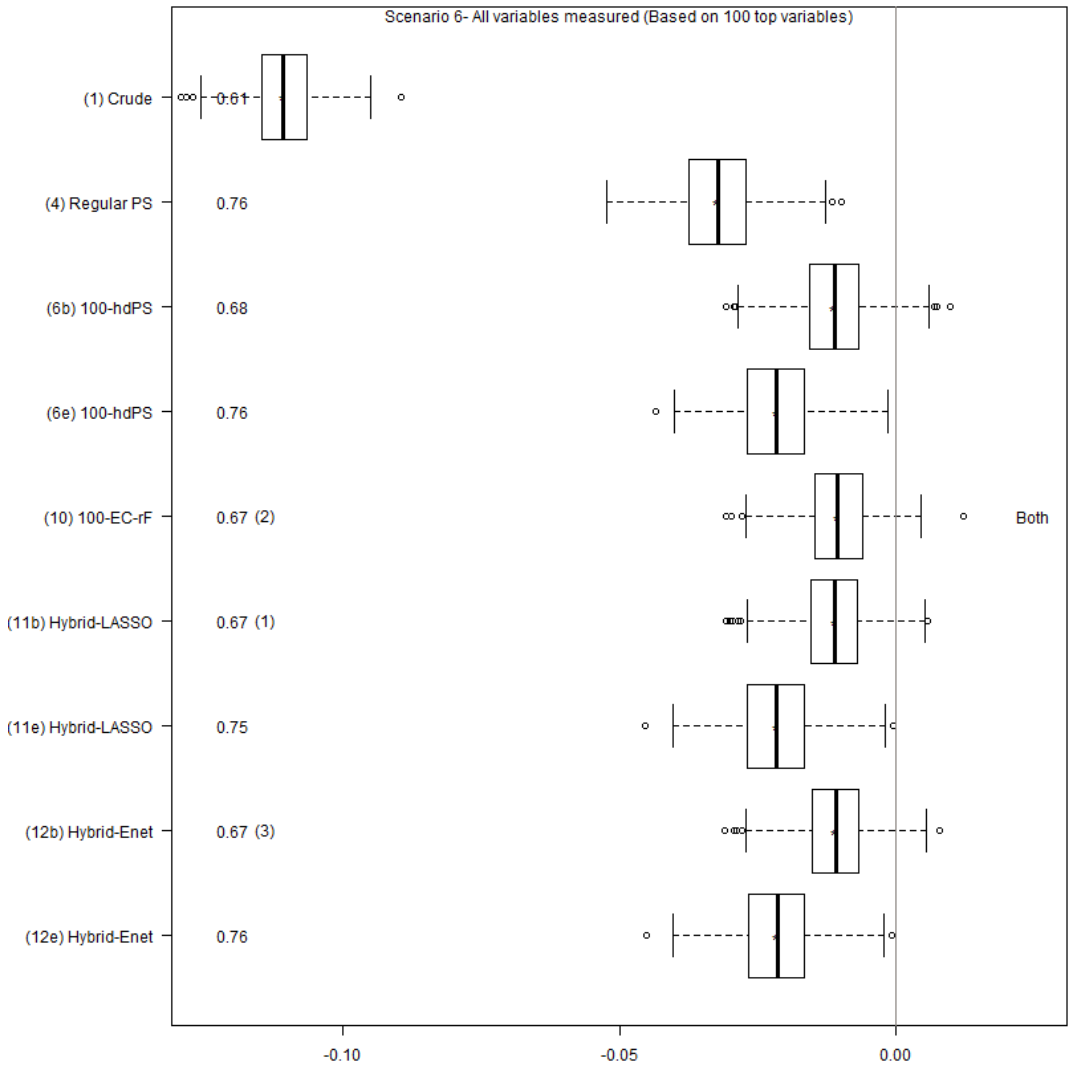
eFigure A.55: Plasmode Simulation Scenario 3-A



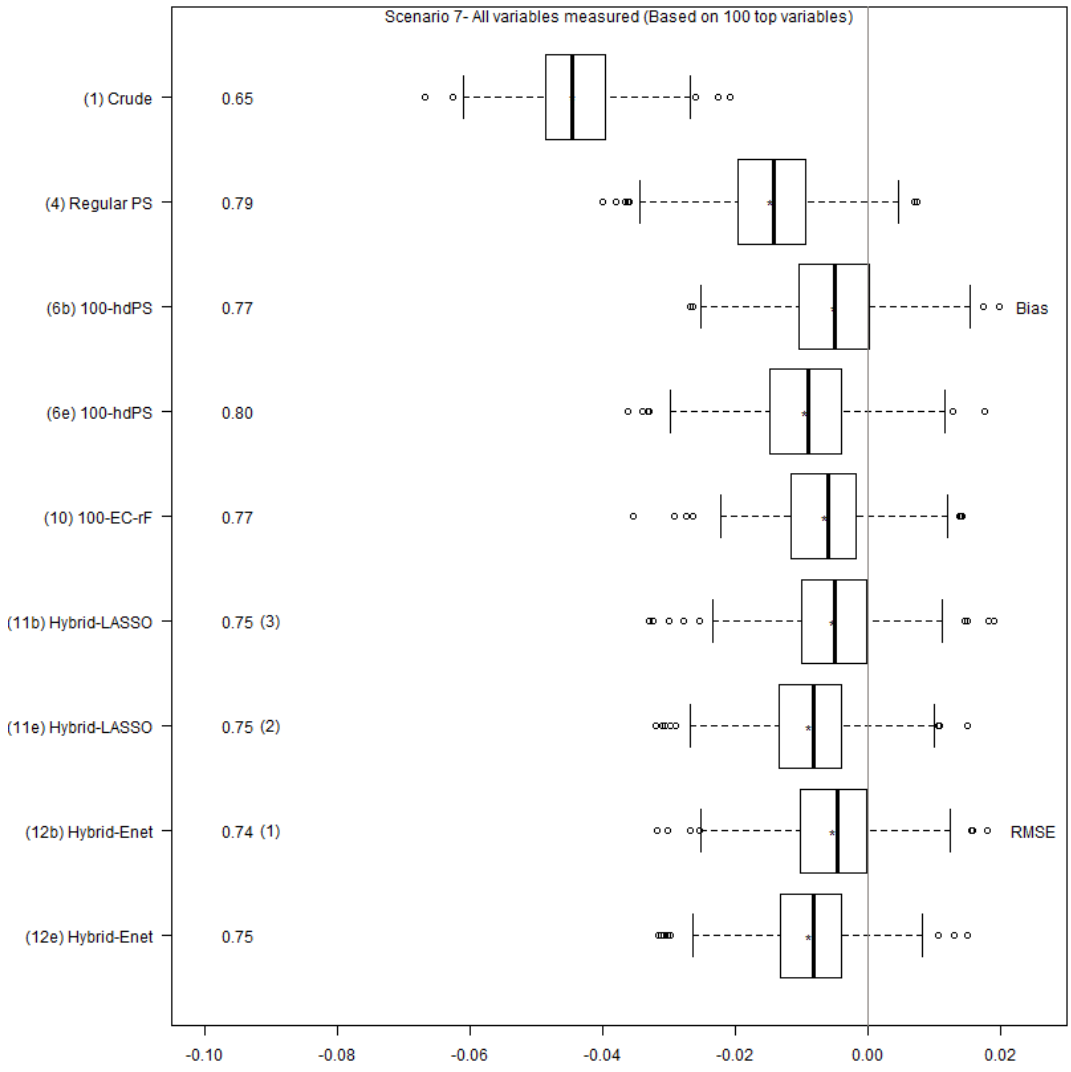
eFigure A.56: Plasmode Simulation Scenario 4-A



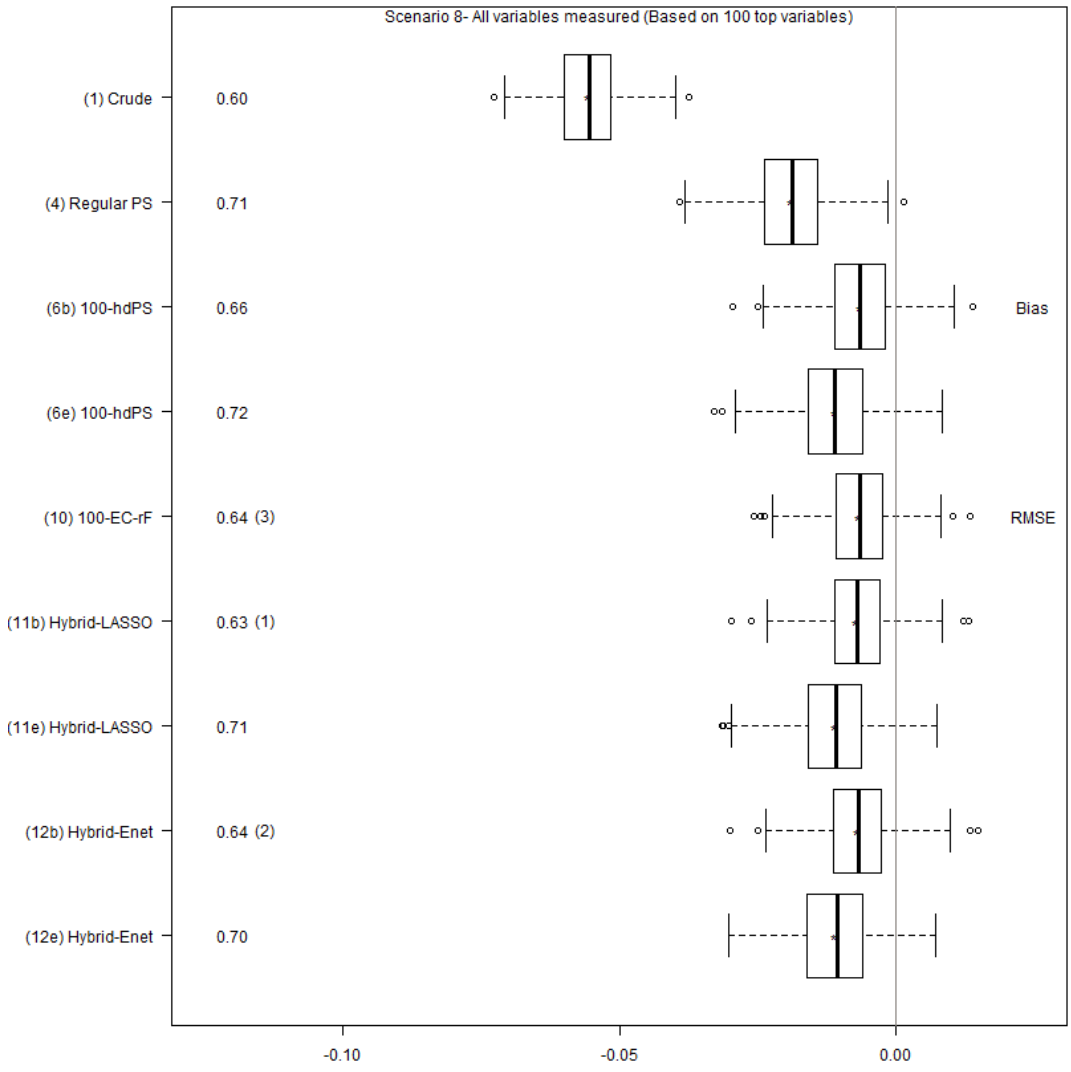
eFigure A.57: Plasmode Simulation Scenario 5-A



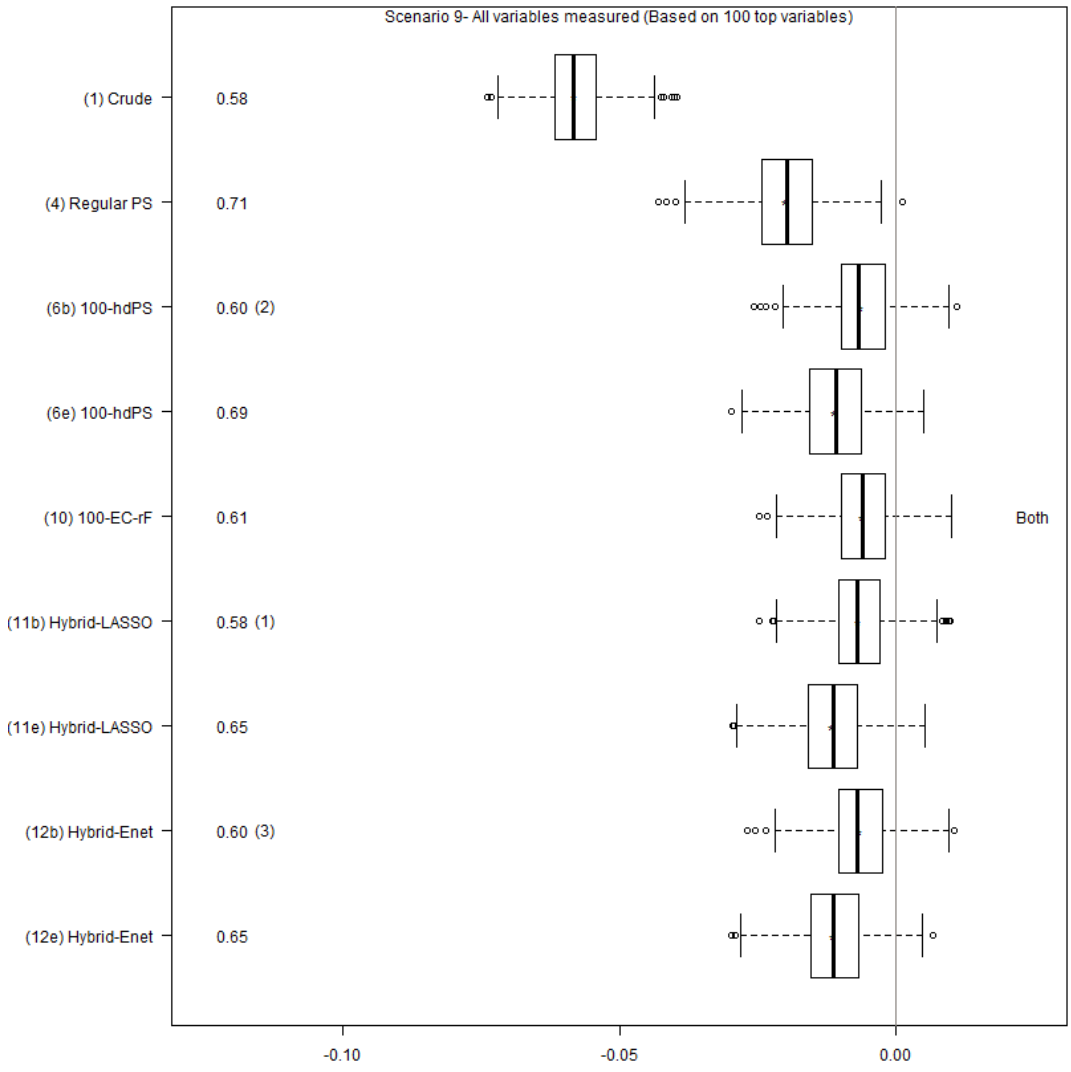
eFigure A.58: Plasmode Simulation Scenario 6-A



eFigure A.59: Plasmode Simulation Scenario 7-A

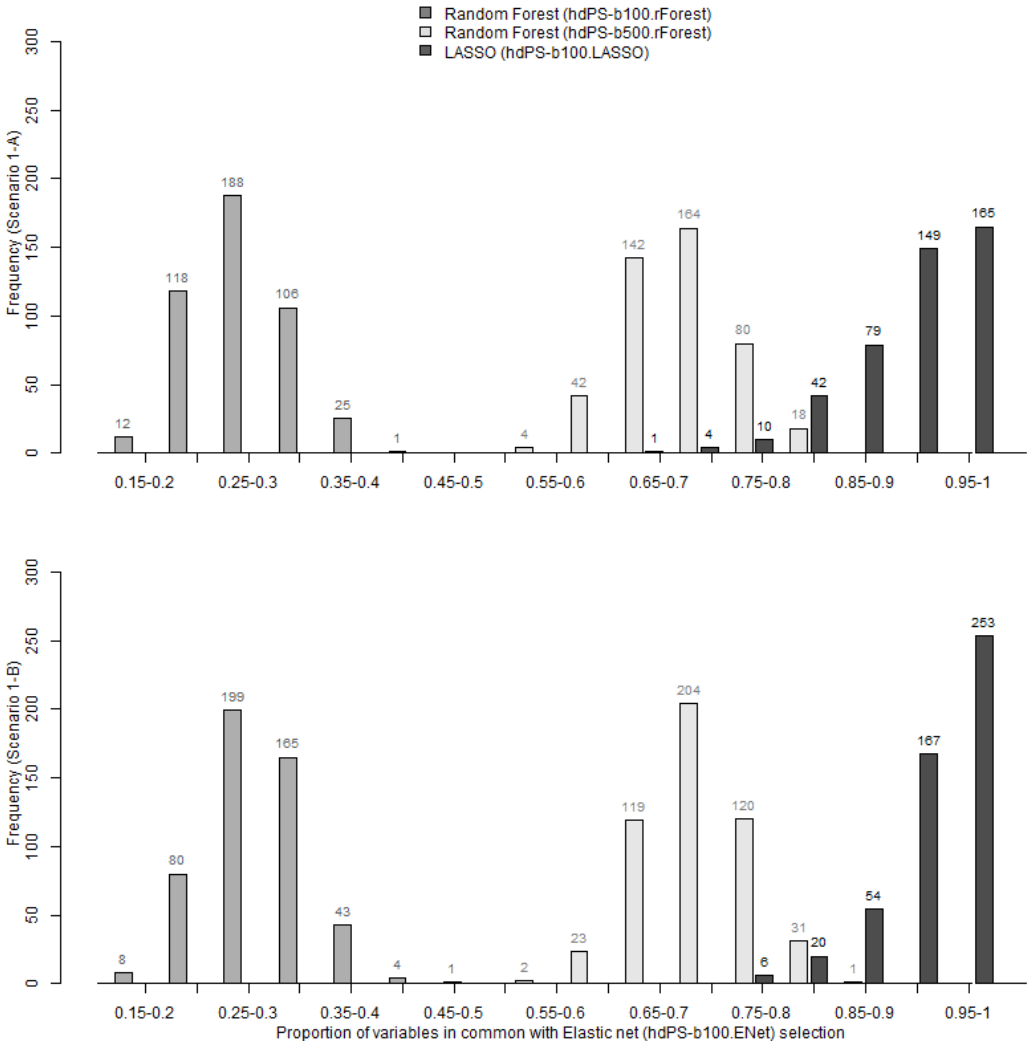


eFigure A.60: Plasmode Simulation Scenario 8-A

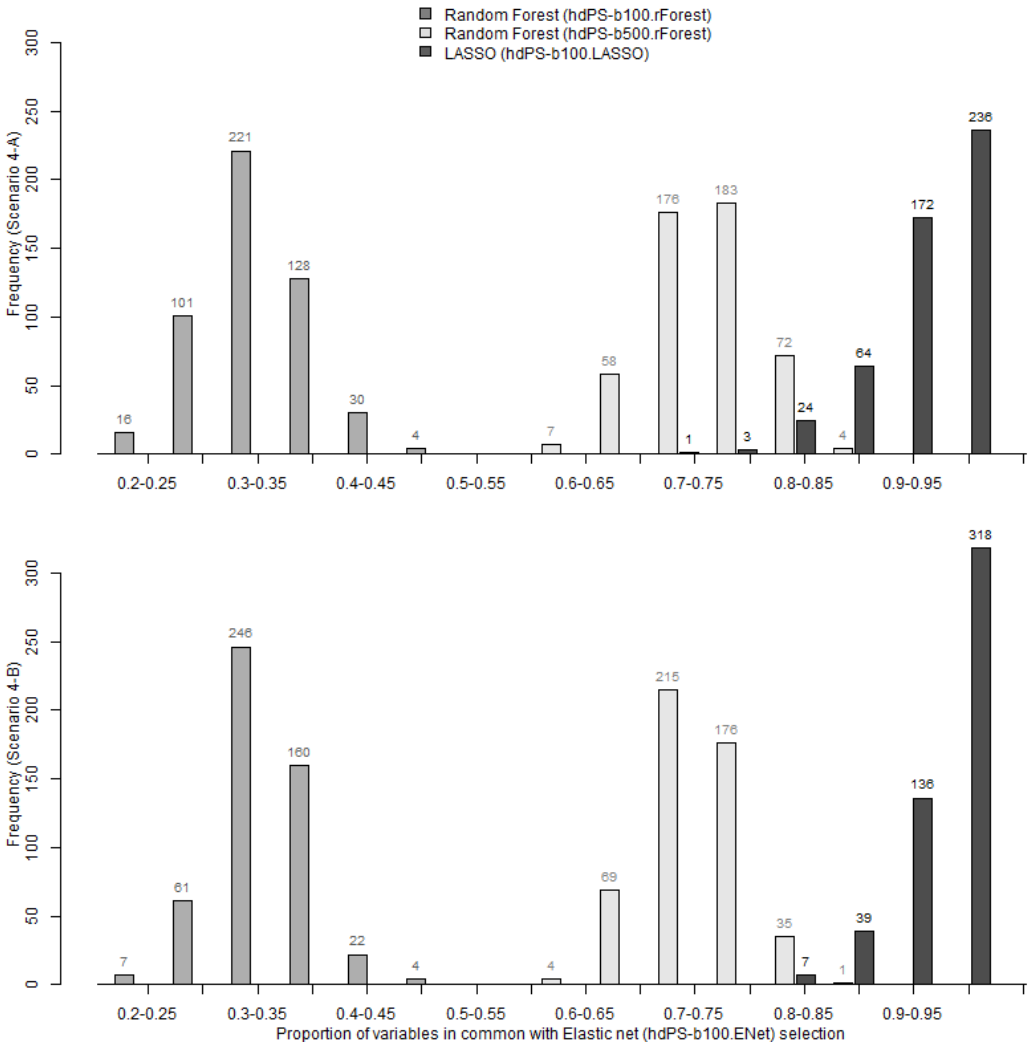


eFigure A.61: Plasmode Simulation Scenario 9-A

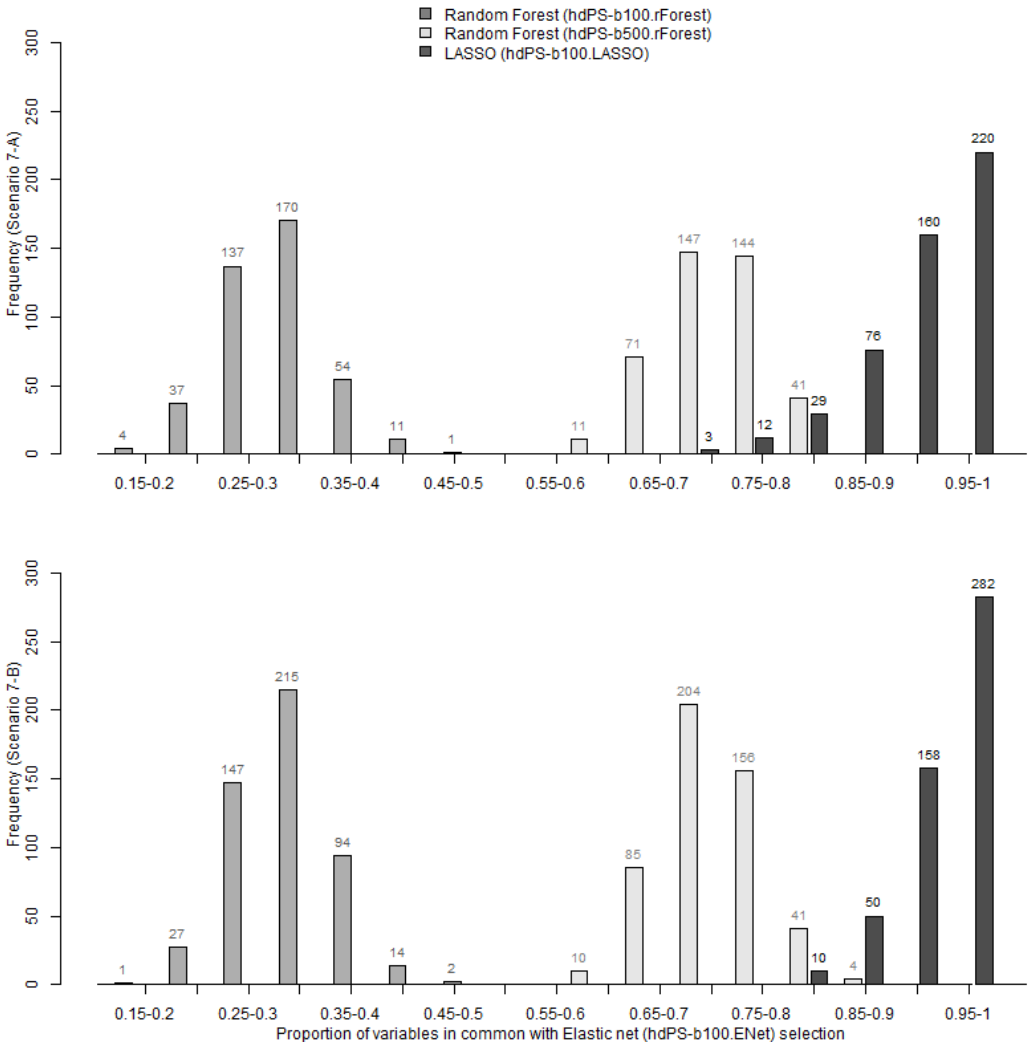
A.10 Proportion of common variables



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.