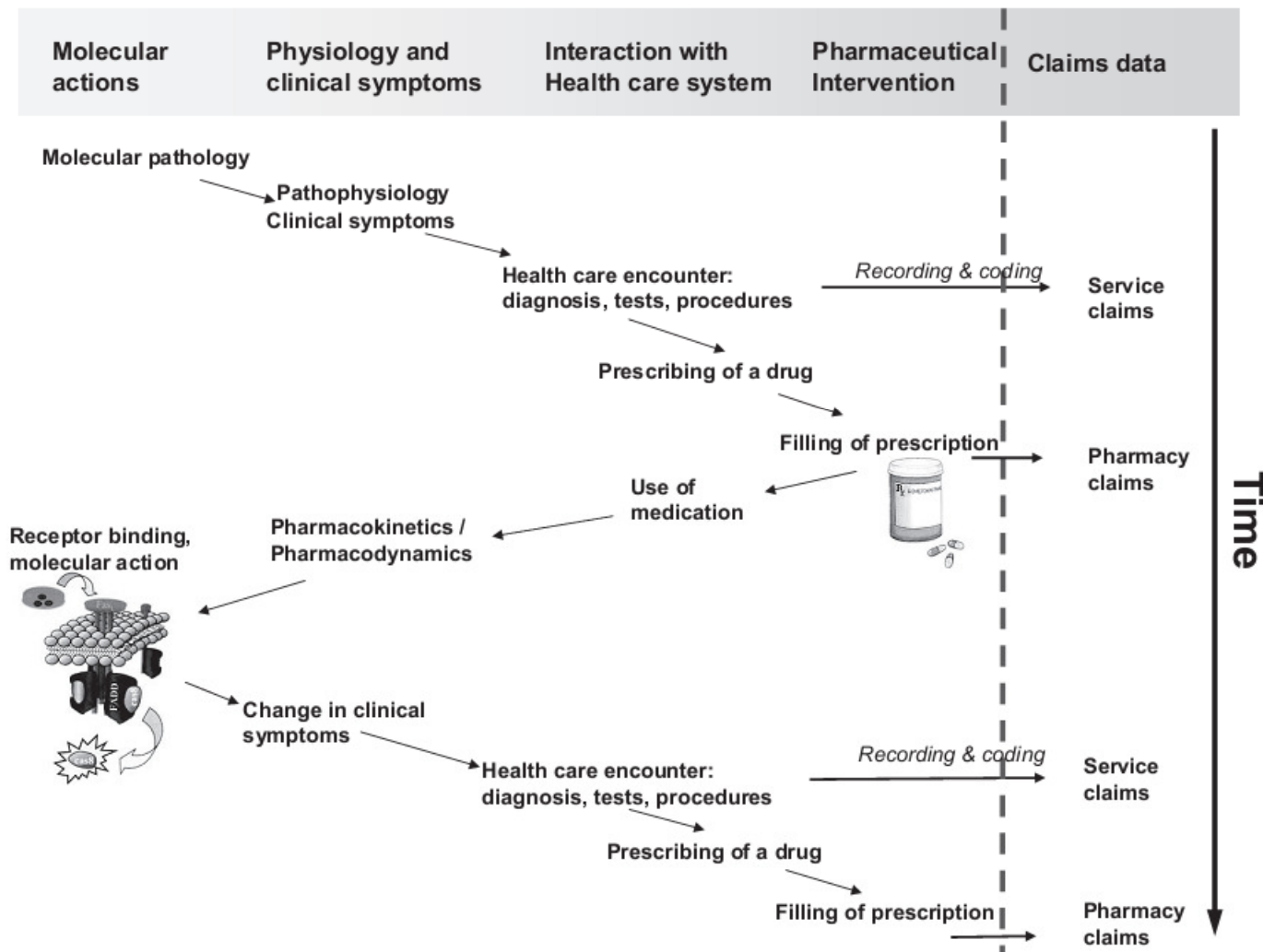
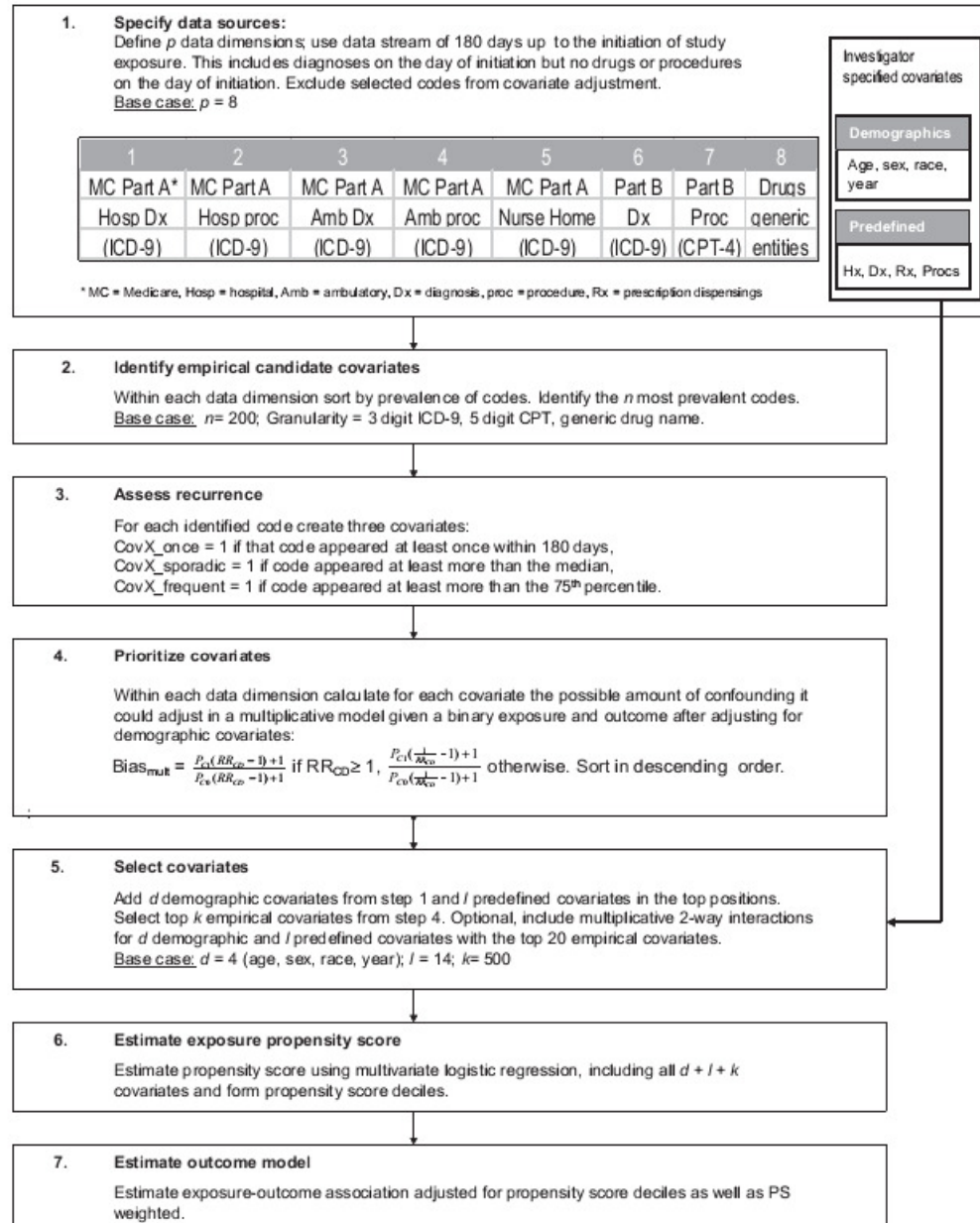


Large Scale Penalized Regression for Propensity Score Estimation in Observational Health Care Data

HDPS*





Ridge logistic regression*

- Maximum likelihood
- and restrictions on coefficients:

$$\sum_{j=1}^P \beta_j^2 \leq s$$

*Hoerl and Kennard (1970).

Lasso*

- Maximum likelihood
- and restrictions on coefficients:

$$\sum_{j=1}^P |\beta_j| \leq s$$

- Does subset selection by shrinking some coefficients to zero

*Tibshirani (1996)

Bayesian Regression*

- Two shrinkage methods
 - Ridge regression - Gaussian prior
$$p(\beta_j | \tau) \sim N(0, \tau)$$
 - Lasso regression - Laplace prior
$$p(\beta_j | \lambda) = \lambda / 2 \exp\{-\lambda |\beta_j|\}$$
- Choosing hyperparameter λ
 - Decide how much to shrink
 - Cross-validation: choose prior to fit left-out data

*Genkin, Lewis, Madigan, (2007) Large-scale Bayesian logistic regression for text categorization -Technometrics

Exposure and outcome

Drug: Sertraline

Outcome: Upper GI Bleeding

Comparator drug: Bupropion

Databases tested: MSLR, MDCR, GE, MD CD, CCAE

Database	Comparator, No outcome	Target, No outcome	Comparator, Outcome	Target, Outcome
<i>MSLR</i>	16743	12599	110	122
<i>MDCR</i>	43267	95123	736	2779
<i>MD CD</i>	130337	146143	1260	1877
<i>GE</i>	115640	200985	300	736
<i>CCAE</i>	611451	486963	4194	3913

MSLR cohorts

MSLR:

	Comparator	Target
Age, mean +/- SD	41.82 +/- 12.84	39.33 +/- 15.24
Male, N (%)	4747 (28.17%)	3381 (26.58%)
Visits, mean +/- SD	7.12 +/- 7.73	7.33 +/- 8.30
Drugs, mean +/- SD	8.27 +/- 6.84	6.91 +/- 6.10
Procedures, mean +/- SD	14.22 +/- 14.51	15.04 +/- 16.40
Charlson index, mean +/- SD	0.73 +/- 1.50	0.69 +/- 1.57

We also consider top 20 procedures, top 20 drugs, and top 20 conditions.

CCAE cohorts

	Comparator	Target
Age, mean +/- SD	42.52 +/- 12.66	39.29 +/- 14.68
Male, N (%)	217322 (35.30%)	152696 (31.11%)
Visits, mean +/- SD	6.79 +/- 7.85	7.10 +/- 8.34
Drugs, mean +/- SD	6.49 +/- 5.85	6.08 +/- 5.76
Procedures, mean +/- SD	10.46 +/- 11.27	11.40 +/- 13.09
Charlson index, mean +/- SD	0.53 +/- 1.23	0.57 +/- 1.40

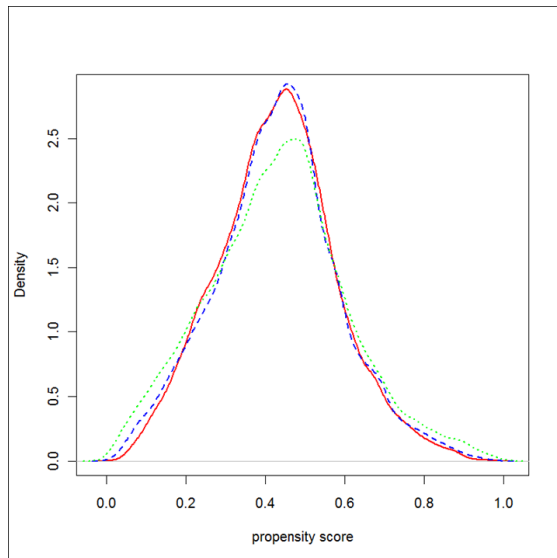
We also consider top 20 procedures, top 20 drugs, and top 20 conditions.

Model for propensity score

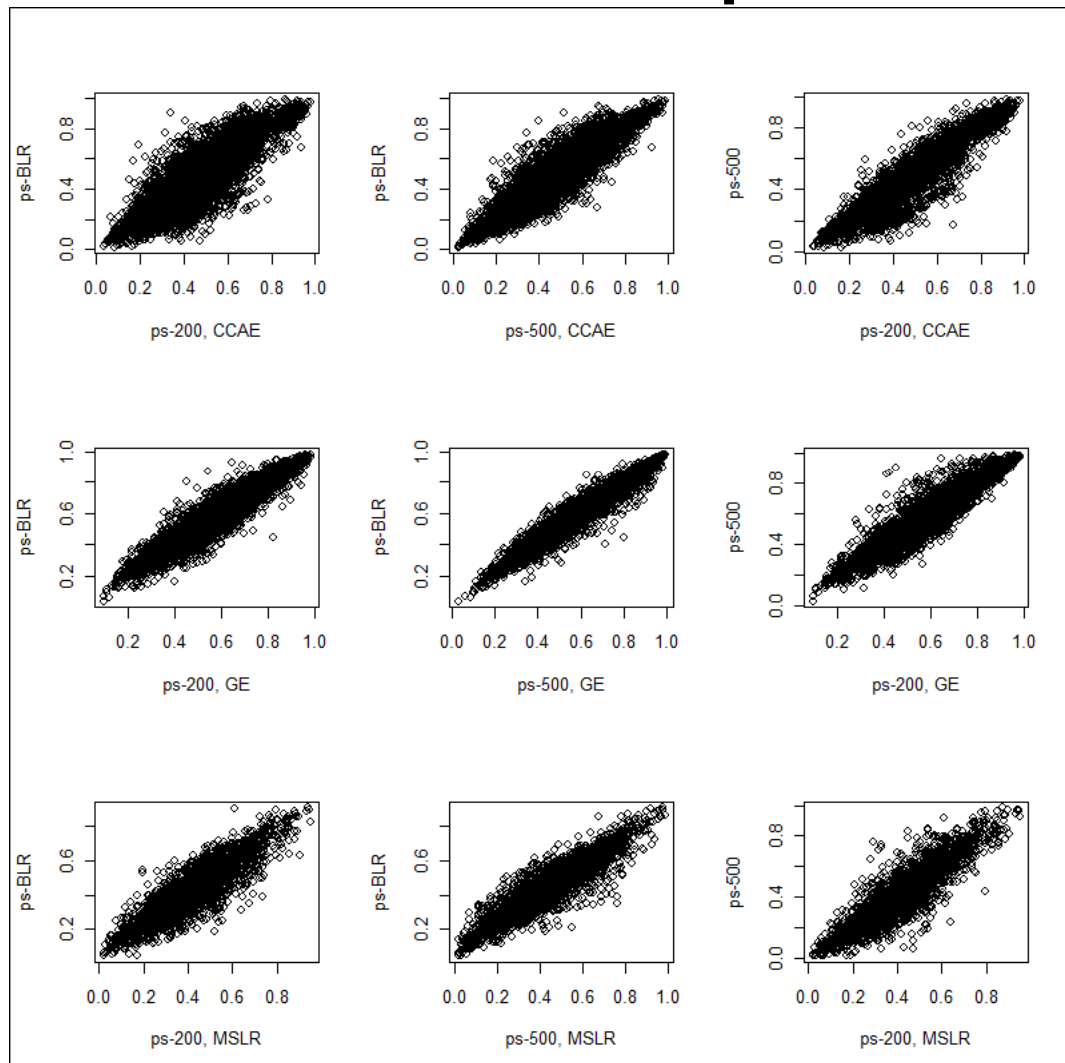
MSLR: includes more than 10000 covariates; all drugs, conditions, procedures, and also indicator variables for gender, age (8 age groups), and number of visits, drugs, conditions, Charlson comorbidity index.

CCA: around 20000 covariates

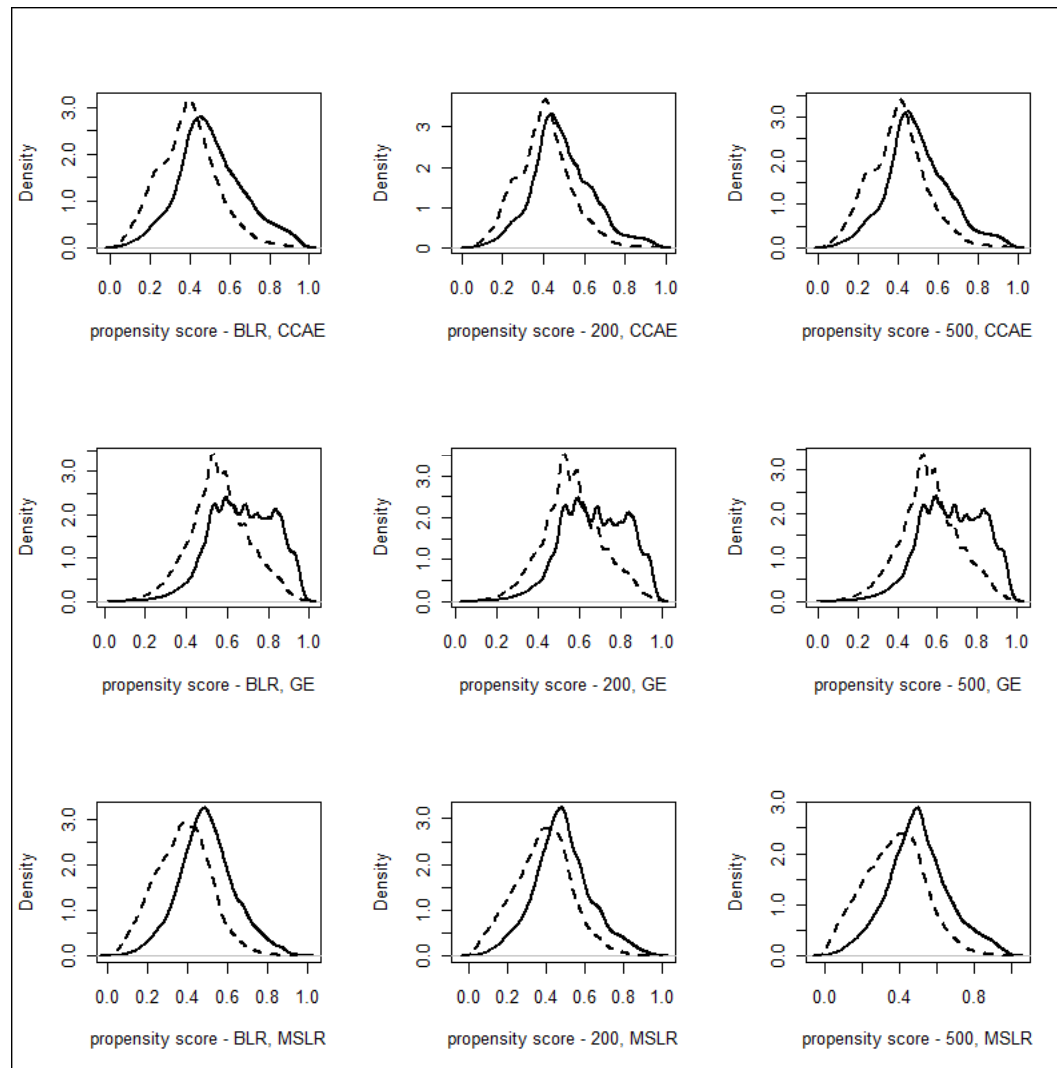
BLR vs. HDPS scatterplots



MSLR: Density estimations of three propensity score methods: red - blr, blue - hdps200, green hdps 500.



Density estimations



Balancing strategies

- Stratification on propensity score
- Matching on propensity score
- Covariate adjustment using propensity score
- Inverse probability of treatment weighting

Assessing Balance

- True propensity score is a balancing score (Rosenbaum, Rubin (1983)).
- Assessing balance: significance testing or standardized differences
- Standardized differences

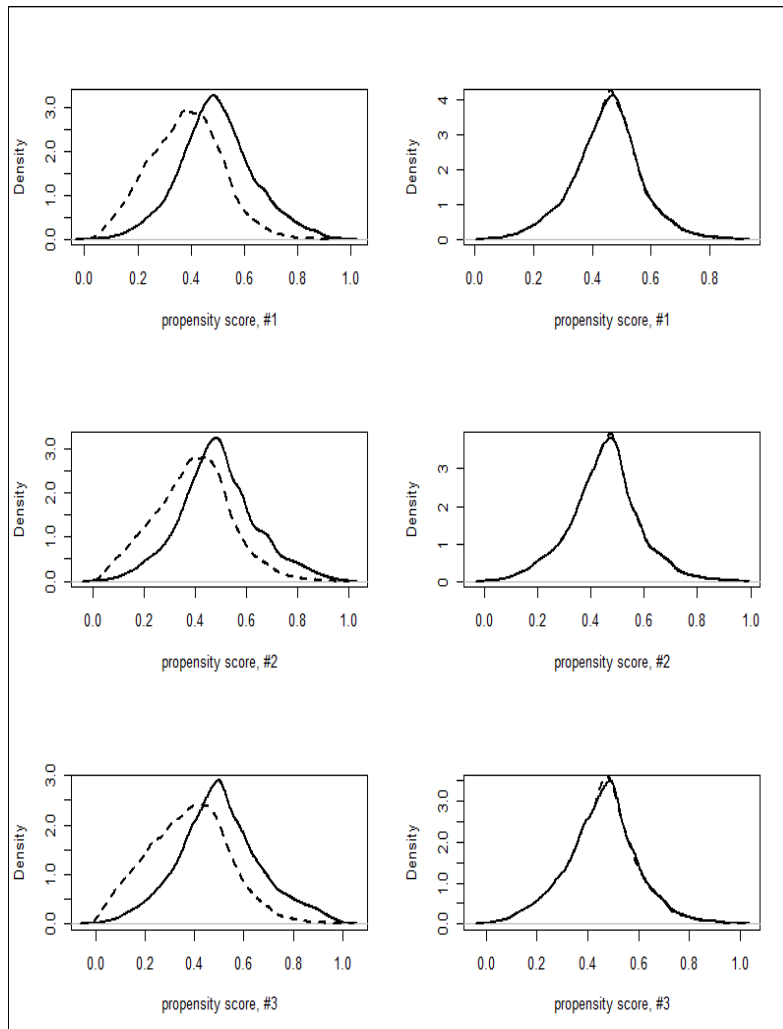
For continuous covariates:

$$d = \frac{|\bar{x}_T - \bar{x}_C|}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$$

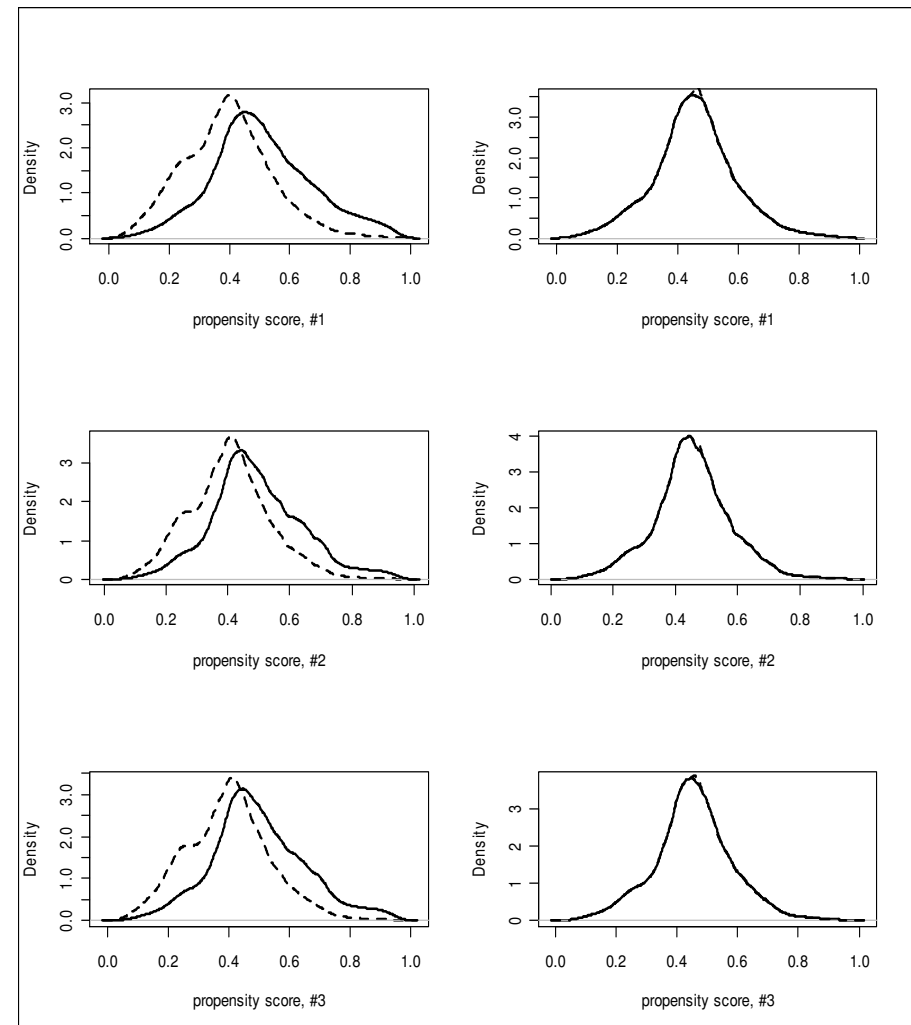
For dichotomous covariates:

$$d = \frac{|\hat{p}_T - \hat{p}_C|}{\sqrt{\frac{\hat{p}_T(1 - \hat{p}_T) + \hat{p}_C(1 - \hat{p}_C)}{2}}}$$

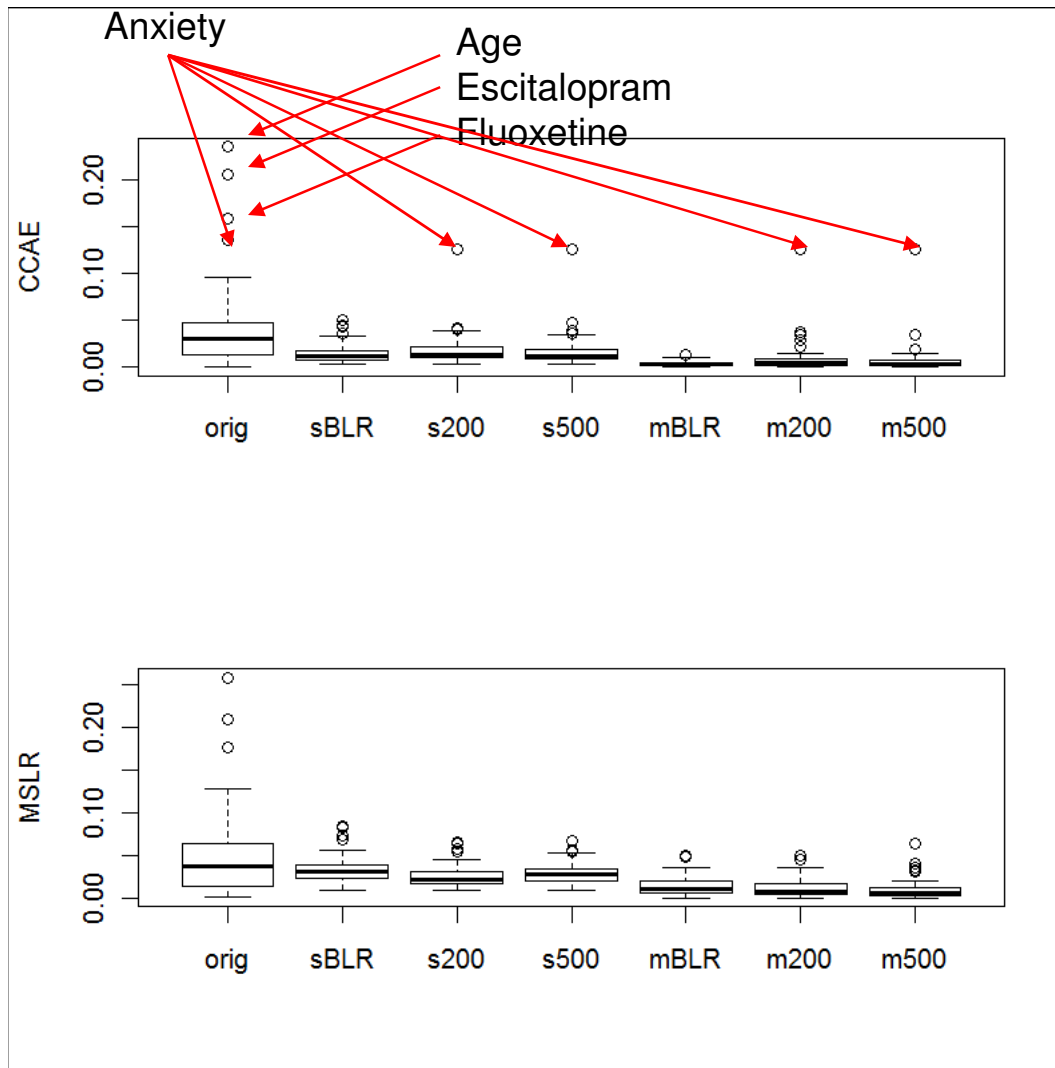
MSLR:



CCAE:

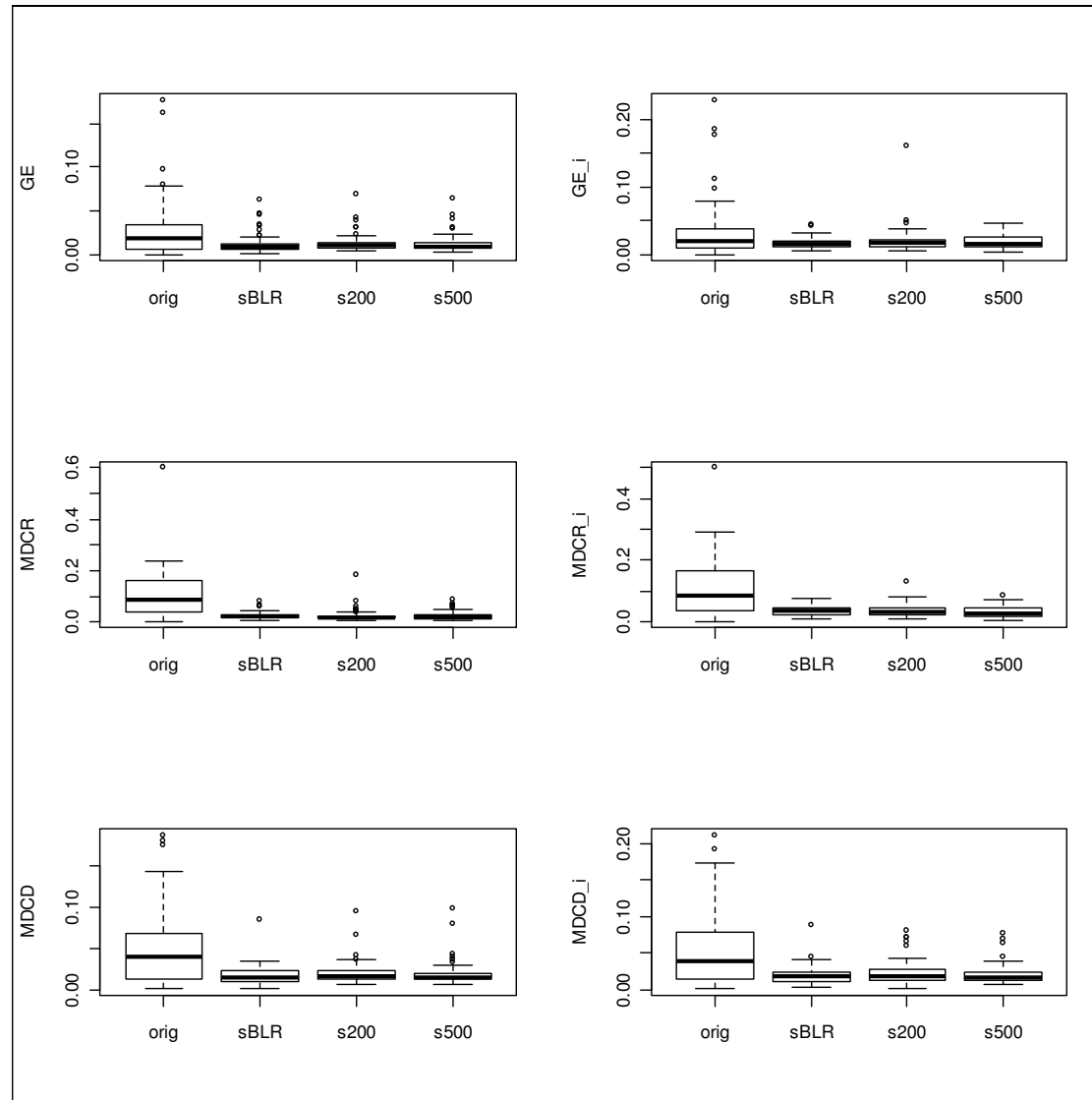


Balancing diagnostics

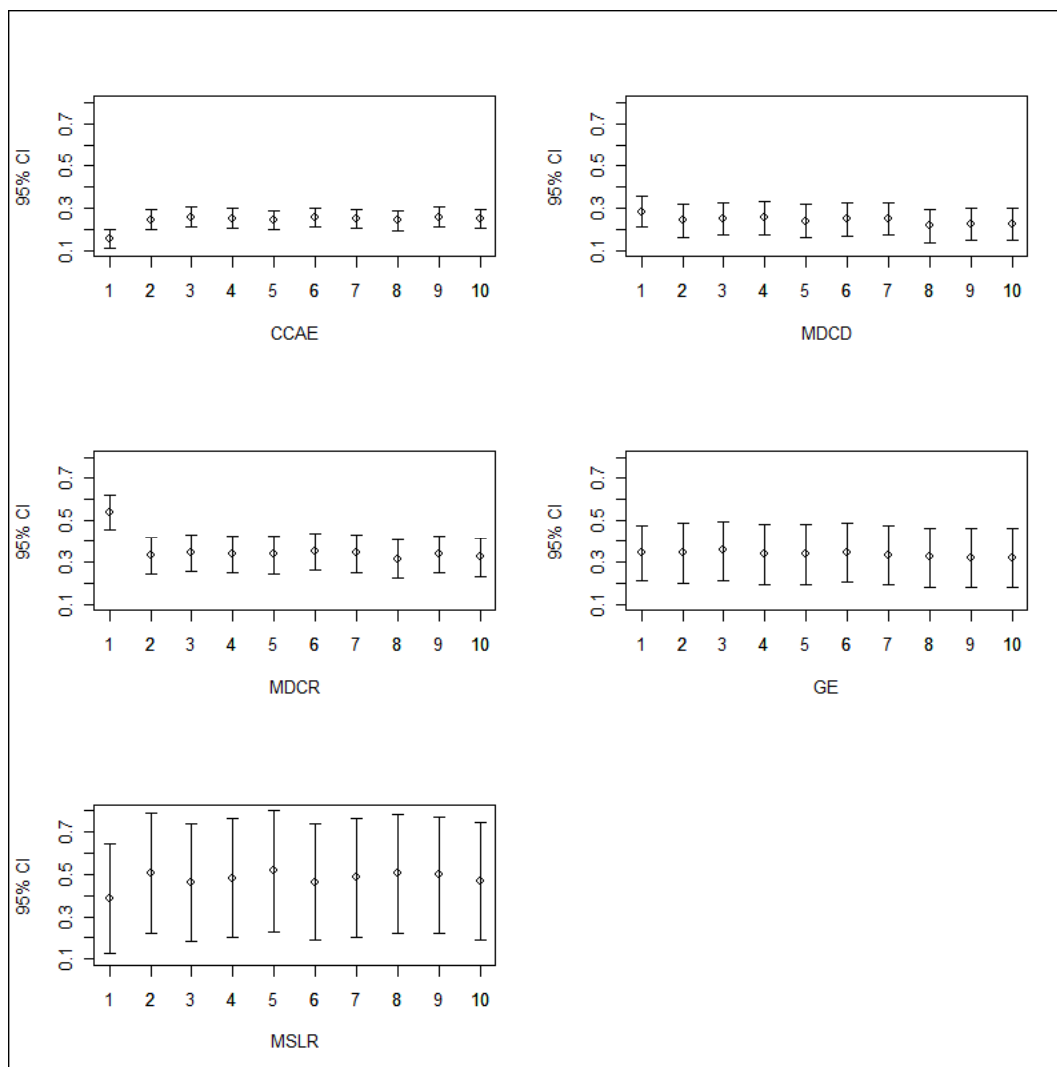


Orig: non-matched data
Stratified data: sBLR, s200, s400
Matched data: mBLR, m200, m500

MORE DATA



Estimates



1: unadjusted, 2: BLR -MH, 3: H200 -MH, 4: H500 - MH, 5: BLR -PS 5 strata, 6: H200 - PS 5 strata, 7: H500 - PS 5 strata, 8: BLR - PS covariate, 9: H200 - PS covariate, 10. H500 - PS covariate.

Conclusions and future steps

- Large scale penalized regression shows similar or superior performance in balancing covariates
- BLR model utilizes all available covariate information and can handle thousands of variables
- Potentially balances on very large number of covariates
- Double robust estimations