



Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study

Journal:	<i>Statistical Methods in Medical Research</i>
Manuscript ID	SMM-19-0433.R2
Manuscript Type:	Original Article
Keywords:	Clinical risk prediction models, Logistic regression, Maximum likelihood, Penalized likelihood, Shrinkage, Firth's correction
Abstract:	<p>When developing risk prediction models on datasets with limited sample size, shrinkage methods are recommended. Earlier studies showed that shrinkage results in better predictive performance on average. This simulation study aimed to investigate the variability of regression shrinkage on predictive performance for a binary outcome. We compared standard maximum likelihood with the following shrinkage methods: uniform shrinkage (likelihood-based and bootstrap-based), penalized maximum likelihood (ridge) methods, LASSO logistic regression, adaptive LASSO, and Firth's correction. In the simulation study, we varied the number of predictors and their strength, the correlation between predictors, the event rate of the outcome, and the events per variable. In terms of results, we focused on the calibration slope. The slope indicates whether risk predictions are too extreme (slope<1) or not extreme enough (slope>1). The results can be summarized into three main findings. First, shrinkage improved calibration slopes on average. Second, the between-sample variability of calibration slopes was often increased relative to maximum likelihood. In contrast to other shrinkage approaches, Firth's correction had a small shrinkage effect but showed low variability. Third, the correlation between the estimated shrinkage and the optimal shrinkage to remove overfitting was typically negative, with Firth's correction as the exception. We conclude that, despite improved performance on average, shrinkage often worked poorly in individual datasets, in particular when it was most needed. The results imply that shrinkage methods do not solve problems associated with small sample size or low number of events per variable.</p>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



1
2
3
4
5
6
7
8
9 **Regression shrinkage methods for clinical prediction models do not guarantee**
10 **improved performance: simulation study**
11
12
13
14
15
16

17 Ben *Van Calster*^{1,2}, Maarten *van Smeden*^{2,3}, Bavo *De Cock*^{1,4}, Ewout W *Steyerberg*²
18
19

20
21
22
23 ¹ KU Leuven, Department of Development and Regeneration, Herestraat 49 box 805, 3000 Leuven,
24 Belgium
25

26
27 ² Department of Biomedical Data Sciences, Leiden University Medical Center, PO Box 9600, 2300 RC
28 Leiden, Netherlands
29

30
31
32 ³ Department of Clinical Epidemiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA
33 Leiden, Netherlands
34

35
36 ⁴ KU Leuven, Department of Accountancy, Finance and Insurance, Naamsestraat 69 box 3525, 3000
37 Leuven, Belgium
38
39

40
41
42
43
44 E-mails ben.vancalster@kuleuven.be, M.van_Smeden@lumc.nl, E.W.Steyerberg@lumc.nl
45
46
47
48

49 Word count of main text: 4134
50
51
52
53
54
55
56
57
58
59
60

Corresponding author: Ben Van Calster, ben.vanecalster@kuleuven.be, +32 16 377788

Abstract

When developing risk prediction models on datasets with limited sample size, shrinkage methods are recommended. Earlier studies showed that shrinkage results in better predictive performance on average. This simulation study aimed to investigate the variability of regression shrinkage on predictive performance for a binary outcome. We compared standard maximum likelihood with the following shrinkage methods: uniform shrinkage (likelihood-based and bootstrap-based), penalized maximum likelihood (ridge) ~~regression~~ methods, LASSO logistic regression, adaptive LASSO, and Firth's correction. In the simulation study, we varied the number of predictors and their strength, the correlation between predictors, the event rate of the outcome, and the events per variable. In terms of results, we focused on the calibration slope. The slope indicates whether risk predictions are too extreme ($\text{slope} < 1$) or not extreme enough ($\text{slope} > 1$). The results can be summarized into three main findings. First, shrinkage improved calibration slopes on average. Second, the between-sample variability of calibration slopes was often increased relative to maximum likelihood. In contrast to other shrinkage approaches, Firth's correction had a small shrinkage effect but showed

1
2
3
4
5
6
7
8
9 low variability. Third, the correlation between the estimated shrinkage and the optimal
10 shrinkage to remove overfitting was typically negative, with Firth's correction as the
11 exception. We conclude that, despite improved performance on average, shrinkage often
12 worked poorly in individual datasets, in particular when it was most needed. The results
13 imply that shrinkage methods do not solve problems associated with small sample size
14 or low number of events per variable.
15
16
17
18
19
20
21
22
23
24
25

26 **Keywords**

27
28
29
30
31
32 Clinical risk prediction models; Firth's correction; logistic regression; maximum
33 likelihood; penalized likelihood; shrinkage
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1. Introduction

When developing clinical prediction models, the ultimate aim is to obtain risk estimates that work well on patients that were not used to develop the model.¹ To do so, we have to keep statistical overfitting under control. Assuming that data collection was done carefully, and according to standardized procedures and definitions, the values in a dataset reflect (1) true underlying distributions of and associations between variables, and (2) some amount of random variability. Overfitting occurs when a prediction model also captures these random idiosyncrasies of the development dataset, which by definition do not generalize to new data from the same population.² The risk of an overfitted model increases when the model building strategy is too ambitious for the available data, for example when the number of variables that are tested as potential model predictors is large given the available sample size.

A well-known rule of thumb for sample size for prediction models is to have at least 10 events per variable (EPV).³⁻⁶ For binary outcomes, the number of events is the number of cases in the smallest of the two outcome levels. ‘Variables’ actually refers to the number of parameters that are considered for inclusion in the model (excluding intercepts). Some parameters may be checked but not included in the final model, and

1
2
3
4
5
6
7
8
9 variables may be modeled using more than one parameter. Recent research has
10 indicated that the $EPV \geq 10$ rule is too simplistic, and highlights that there are no good
11 rules of thumb regarding sample size.⁷⁻¹¹ Therefore, the use of shrinkage methods is
12 recommended when sample size is small.^{5,6} Several studies have suggested that model
13 performance improves on average when shrinkage methods are applied.^{5,9,12-17} Some
14 have suggested that shrinkage may be needed for EPV values up to 20 if the model is
15 prespecified.¹ When variable selection has to be performed to develop the model, the
16 required EPV for reliable selection may increase to 50.¹
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 Most regression shrinkage methods deliberately induce bias in the coefficient estimates,
32 by shrinking them towards zero, in order to reduce the expected variance in the
33 predictions. As a consequence, for models with a binary outcome, these methods aim to
34 prevent predicted risks that are too extreme, i.e. where small risks are underestimated,
35 and high risks overestimated. This leads to better expected mean squared error of the
36 predictions.¹⁸ Since prediction focuses on reliable predictions, inducing bias in the
37 model coefficients is not a key concern. Therefore, it seems that the use of shrinkage
38 methods is always good when sample size is limited. Moreover, standard maximum
39 likelihood estimation suffers from small sample bias leading to exaggerated coefficient
40 estimates (i.e. away from zero).^{6,19} However, some observations are puzzling. Hans van
41 Houwelingen already noted that ‘it is surprising to observe that the estimated shrinkage
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 factors can be quite off the mark and are negatively correlated with optimal shrinkage
10 factor'.²⁰ This would imply that shrinkage methods shrink too little when it is really
11 needed, and vice versa. However, van Houwelingen's paper included only small
12 simulation study focusing on uniform shrinkage factors. It is of interest to see whether
13 this also occurs with ~~more modern~~ approaches to regression shrinkage, such as
14 LASSO, ridge, and Firth's correction.^{19,21,22} Other studies suggest that some methods
15 result in too much shrinkage on average, as indicated by an average calibration slope
16 larger than one.^{9,14,16,23} In Box 1, we present an illustration dealing with a prediction
17 model for ovarian cancer diagnosis,²⁴ to illustrate that standard regression and
18 regression shrinkage may be more variable in performance than many would think.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

36 The aim of this simulation study was to investigate the performance of various modern
37 shrinkage approaches for the validity of clinical prediction models that are developed
38 with small number of predictors relative to the total sample size (low dimensional). This
39 implies a situation in which some preselection of potentially important predictors has
40 been done before the modeling (e.g. by expert opinion or based on previous studies).
41 We address the performance on average, as well as performance for individual
42 simulation runs. The latter is done by evaluating the between-sample variability in the
43 amount of shrinkage provided by various methods, and the correlation between
44 estimated shrinkage and optimal shrinkage.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

2. Materials and methods

2.1. Regression methods

We will apply standard logistic regression based on maximum likelihood estimation, and compare this to a collection of shrinkage methods within the context of logistic regression. We apply likelihood-based and bootstrap-based uniform shrinkage methods,^{12,25} methods that directly shrink coefficient estimates without or with variable selection (~~classical ridge logistic regression and a more general penalized maximum likelihood method~~),^{21,22,26-28} ~~methods that directly shrink coefficient estimates with selection~~ (least absolute shrinkage and selection operator (LASSO) and ~~adaptive LASSO~~),^{22,29} and Firth's penalized likelihood.^{19,30} We will discuss each method in what follows.

Standard logistic regression. This is the reference method, in which coefficients are determined by maximum likelihood (ML). Hence, no shrinkage is applied here. When the outcome variably Y equals 1 for an event and 0 for a non-event, the probability of an

event ($Y = 1$) for patient i (π_i) is estimated based on a weighted combination of p predictor variables X_j . We define π_i as $P(Y = 1|\mathbf{x}_i)$, with $i = 1, \dots, n$, and $\mathbf{x}_i = (1, x_{1,i}, \dots, x_{p,i})'$. Assuming only linear effects and no interactions between the predictors, the logistic regression has the following form:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \sum_{j=1}^p \beta_j x_{ij} = \mathbf{x}_i' \boldsymbol{\beta},$$

where $\pi_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}}$, and $\boldsymbol{\beta}$ a column vector containing the intercept α and the coefficients β_j . Coefficient estimates $\hat{\alpha}$ and $\hat{\beta}_j$ are obtained by finding the maximum of the log-likelihood function:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \{y_i \log(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \log(1 - \pi_i(\boldsymbol{\beta}))\}.$$

Likelihood-based uniform shrinkage (LU). This method uses the likelihood-ratio statistic to compute a uniform shrinkage factor

$$S_{LU} = \frac{\chi_{model}^2 - df}{\chi_{model}^2},$$

where χ_{model}^2 is the likelihood-ratio statistic of the fitted model based on standard maximum likelihood and df is the degrees of freedom for the number of candidate predictors considered for the model.²⁵ The shrunk model coefficients are then calculated

1
2
3
4
5
6
7
8
9 as $\hat{\beta}_{j,LU} = s_{LU}\hat{\beta}_j$. After adjusting the coefficients, we re-estimated the intercept to
10
11 guarantee that the average predicted risk equaled the event rate.
12
13

14
15
16
17 *Bootstrap-based uniform shrinkage (BU)*. The uniform shrinkage factor s can also be
18
19 computed using a bootstrap procedure:¹²
20
21

- 22
23
24
25
26 1. A bootstrap sample is taken from the original data sample, that is, a random
27
28 sample with replacement of the same size as the original sample.
- 29
30
31
32 2. If a selection procedure was used to select variables this is also applied in the
33
34 bootstrap samples. The regression coefficients are estimated again on the bootstrap
35
36 sample, $\hat{\beta}_{bt}$.
- 37
38
39
40 3. The linear predictor for each of the observations in the original sample is
41
42 calculated using $\hat{\beta}_{bt}$.
- 43
44
45
46 4. In the original sample, the linear predictor obtained in the previous step is used to
47
48 predict the outcome using maximum likelihood. Retain the coefficient for the
49
50 regression of the linear predictor.
- 51
52
53
54 5. Repeat the procedure, steps one to four, and the average coefficient from step four
55
56 provides the shrinkage factor s_{BU} . We used 200 repetitions.
57
58
59
60

6. The shrunk coefficients are calculated as $\hat{\beta}_{j,BU} = s_{BU}\hat{\beta}_j$.

7. Re-estimate the intercept using maximum likelihood while keeping $\hat{\beta}_{j,BU}$ fixed.

Classical ridge logistic regression. Regression shrinkage is implemented via the ridge penalty, also known as the quadratic or L2-penalty.²¹ Ridge regression was extended to logistic regression initially by Schaefer and colleagues, and later by Le Cessie and Van Houwelingen.^{26,27} The following penalized version of the log-likelihood function is maximized:

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p \beta_j^2.$$

The tuning parameter, λ , controls the amount of shrinkage. ~~The optimal value for this parameter can be estimated by, for example, generalized cross validation (GCV).~~ Ridge regression shrinks the estimated coefficients towards zero (on average), with higher values of λ leading to more shrinkage. This implicitly induces bias in the coefficients. Note that coefficients will not be shrunk to zero and that the intercept term is not penalized. The shrinkage parameter λ is a hyperparameter that has to be estimated ('tuned'). We used 10-fold cross-validation to find the value for λ that minimized the deviance, using a grid of 251 possible values between zero (no shrinkage) and 64 (very large shrinkage). The 250 non-null values were equidistant on logarithmic scale. We used the `glmnet` R package to implement ridge logistic regression.³²

1
2
3
4
5
6
7
8
9
10
11
12 *General penalized maximum likelihood estimation.* Ridge logistic regression is a special
13 case of Harrell describes a penalized maximum likelihood (PML) that maximizes the
14 following function: ~~estimation procedure that maximizes a penalized version of $\ell(\boldsymbol{\beta})$ as~~
15 ~~a more general method than ridge.~~²⁸ The following function is maximized:

$$\ell(\boldsymbol{\beta}) - 0.5\lambda\sum_{j=1}^p (s_j\beta_j)^2,$$

21
22
23
24
25 where s_j are scaling factors that allow more flexibility than classical ridge. In our study,
26 will apply the method as suggested by Harrell.²⁸ We set the scale factors to be the
27 standard deviation of the predictor. As our predictors are simulated as standard normal
28 variable, and we standardize the variables before fitting models, this approach does not
29 differ from classical ridge. However, Harrell suggests to tune the shrinkage parameter
30 based on a Akaike Information Criterion instead of cross-validation, because it is faster
31 and performs slightly better.²⁸ Following Harrell's suggestion, the tuning parameter was
32 chosen using the corrected Akaike Information Criterion using a similar grid as for
33 classical ridge.^{28,33} The rms R package was used to implement this method. In tables
34 and figures, we refer to this method with the abbreviation PML, and to classical ridge
35 regression with the abbreviation L2.

Classical LASSO logistic regression. LASSO is similar to ridge, but uses the L1-penalty that poses a constraint on the sum of the absolute value of the estimated coefficients.²²

For logistic regression, the LASSO optimizes the following function:

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p |\beta_j|.$$

The L1-penalty allows coefficients to be shrunk to zero, and hence LASSO performs variables selection as well. The shrinkage parameter was tuned using cross-validation in the same way as for classical ridge logistic regression. The glmnet R package was used.

Adaptive LASSO (AL). The Adaptive LASSO is a variant of the LASSO where a weight is given for each parameter in the penalty term, in order to obtain variable selection consistency.²⁹ The optimized function is:

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p w_j |\beta_j|,$$

where $w_j = \frac{1}{|\hat{\beta}_j^{init}|^\gamma}$ contains adaptive weights. The $\hat{\beta}_j^{init}$ are initial coefficient estimates

for the predictors. We used the maximum likelihood estimate $\hat{\beta}_j$ as $\hat{\beta}_j^{init}$, and fixed γ at unity.^{15,29} Adaptive LASSO shrinks higher absolute values of $\hat{\beta}_j^{init}$ less than lower

values. We tuned the shrinkage parameter using cross-validation as for classical LASSO. The glmnet R package was used.

1
2
3
4
5
6
7
8
9
10
11
12 *Firth's penalized likelihood.* Firth developed a procedure to remove the first order bias
13
14 in the regression coefficients based on maximum likelihood.^{19,30} To do so, modified
15
16 score functions are used to estimate model coefficients. This avoids problems with
17
18 separation, but also shrinks the coefficients. In addition, Firth's correction reduces the
19
20 variance. In terms of the log-likelihood, Firth's correction optimizes

$$\ell(\boldsymbol{\beta}) + 0.5 \log |I(\boldsymbol{\beta})|,$$

26
27 where $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. We used the `logistf` R
28
29 package to implement this method. For making predictions based on Firth's correction,
30
31 the intercept has to be corrected.¹⁶ We used the same intercept re-estimation procedure
32
33 as for the uniform shrinkage methods.
34
35
36
37
38
39

40 2.2. Simulation setup

41
42
43
44
45 We simulated data to predict a binary outcome. We used a full factorial simulation setup
46
47 varying the following factors: EPV, the number and strength of predictors, the
48
49 correlation between predictors, and the outcome event rate (Table 1). In total, this gave
50
51 us 60 simulation scenarios. In the setting with five true predictors, the true coefficients
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 of the predictors were 0.2, 0.2, 0.2, 0.5, and 0.8. These values were based on the
10 Cohen's d measure of effect size, and would correspond to having three weak predictors
11 (odds ratio 1.22), one moderate predictor (odds ratio 1.65), and one strong predictor
12 (odds ratio 2.23).³¹ In the setting with 10 true predictors, six had a coefficient of 0.2,
13 two had a coefficient of 0.5 and two had a coefficient of 0.8. Noise predictors had
14 coefficients of 0. The chosen values of the simulation factors had an impact on the true
15 c-statistic (i.e. area under the receiver operating characteristic curve) of the model, the
16 sample size of the simulated datasets, and the number of cases with an event (Table 2).
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 For every scenario, the simulations were performed as follows. First, for each of 1 000
32 000 individuals the predictor values were generated by draws from a standard
33 multivariate normal distribution, with equal pairwise correlations. The true model
34 formula (linear predictor) was applied to each patient, with the intercept chosen to
35 obtain the target event rate (Table 2). The inverse logit of the linear predictor was the
36 true risk for that individual. Then, the outcome for each patient was generated through a
37 Bernoulli trial using the true risk. A different dataset, but also with 1 000 000
38 individuals, was generated for model validation. Predictors and outcomes were
39 generated analogous to the development population, which means that our out-of-
40 sample performance corresponds to a large sample internal validation setting.
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12 We executed 1 000 simulation runs per simulation condition. For each run, we
13
14 generated a development dataset of the appropriate size (Table 2) by randomly drawing
15
16 without replacement from the development population. The event rate was fixed at the
17
18 target value in each development dataset by applying stratified sampling. Next, the
19
20 predictor variables were standardized, and all types of models were fitted. ~~Ridge,~~
21
22 ~~LASSO and adaptive LASSO models were estimated using the glmnet R package.³²~~
23
24 ~~PML models were developed using the rms R package.²⁸ For ridge and (adaptive)~~
25
26 ~~LASSO, the tuning parameter was selected from a grid of 251 values between zero (no~~
27
28 ~~shrinkage) and 64 (very large shrinkage). The 250 non-null values were equidistant on~~
29
30 ~~logarithmic scale. We used 10-fold cross-validation that minimized the cross-validated~~
31
32 ~~deviance. Following Harrell's suggestion, the tuning parameter for PML was chosen~~
33
34 ~~using the corrected Akaike Information Criterion using a similar grid.^{28,33} Firth's~~
35
36 ~~penalized models were developed using the logistf R package. When using standard~~
37
38 maximum likelihood, separation was suggested when R warned for fitted probabilities
39
40 of zero or one, or when the model did not converge. In these circumstances, results for
41
42 standard maximum likelihood were replaced with results based on Firth's correction,
43
44 because this is a situation where the use of the method is indicated.^{19,30} For LU and BU,
45
46 the shrinkage factor s was calculated for the model using Firth's correction (with
47
48 bootstrap models for BU also based on Firth's correction). ~~The Harrell's suggested~~
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 PML method often resulted in an error when there was the suggestion of separation. In
10 these cases, we used Firth's correction instead of the PML algorithm. In this way, we
11 could avoid the exclusion of samples that were suggestive of separation.³⁴ For logistic
12 regression with bootstrap uniform shrinkage, bootstrap models suggestive of separation
13 were replaced by other bootstrap replicates without separation.
14
15
16
17
18
19
20
21
22
23

24 The resulting models were validated on the accompanying full validation dataset. We
25 calculated the c-statistic and the calibration slope. Because the development and
26 validation data are based on identical populations, the calibration intercept was of little
27 interest and therefore not calculated.³⁵ At internal validation (i.e. when the underlying
28 population is the same), the calibration slope measures bias of risk predictions in terms
29 of spread.^{35,36} A slope below unity suggests that predictions are too extreme: low risks
30 are underestimated, high risks are overestimated. A slope above unity suggests the
31 opposite. We calculated median slopes to assess the deviation from the target value of
32 unity. To investigate the variability in the slope, we calculated the median absolute
33 deviation (MAD) of the log(slope). To combine bias (deviation of slope from unity on
34 average), and variability, we calculated root mean squared distance from the target
35 value (RMSD) of the log(slope) over the 1 000 runs. We used the logarithm of the slope
36 to acknowledge its asymmetry. A slope of 0.5 (half the target) corresponds to a similar
37 quantitative deviation to a slope of two (double the target), but in opposite directions.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 The RMSD was calculated as the square root of the mean of $(\log(1) - \log(\text{slope}))^2$
10 over the 1 000 runs. Finally, we calculated the Spearman correlation between the
11 estimated shrinkage and the optimal shrinkage over the 1 000 simulation runs. The
12 optimal shrinkage was defined as $\log(1) - \log(\text{slope}_{\text{ML}})$, with slope_{ML} the slope for the
13 standard maximum likelihood model. The estimated shrinkage for a specific shrinkage
14 approach was defined as $\log(\text{slope}_{\text{shrinkage}}) - \log(\text{slope}_{\text{ML}})$. To calculate MAD,
15 RMSD, and correlations, we winsorized slopes at 0.01 to avoid problems with rare
16 instances of negative calibration slopes. When no variables were selected by (adaptive)
17 LASSO, the calibration slope was arbitrarily set at 10 to reflect the extreme amount of
18 underfitting.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

36 R code used for the simulations can be found at GitHub
37 (<https://github.com/benvancalster/shrinkagesim/>).
38
39
40
41
42
43

44 3. Results

45
46
47
48
49

50 There were few runs where separation was suggested (Table S1), except in the scenario
51 with three EPV, 10 true predictors, 0.5 correlation and 0.5 event rate. Generally, results
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 differed little between the five predictor and 10 prediction scenarios, therefore we focus
10 here on the scenarios with five true predictors for the main document. Detailed results
11 here on the scenarios with five true predictors for the main document. Detailed results
12 for all scenarios are provided in supplementary tables and figures.
13
14
15
16
17
18
19

20 3.1. Performance on average 21 22 23 24

25 The median calibration slope approached unity for all methods as EPV increased
26 (Figure 1, Figure S1, Table S2). The standard maximum likelihood model yielded the
27 lowest median calibration slopes. For classical ridge regression, the median slope at
28 lower EPV values was consistently above unity, suggesting too much shrinkage on
29 average. [Harrell's](#) PML and LASSO were better, but in many scenarios showed median
30 slopes above unity as well. Other methods generally had median slopes below unity,
31 with bootstrap uniform shrinkage usually having median slopes closest to unity. The use
32 of Firth's correction was slightly better than maximum likelihood.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 The average c-statistics also converged to their respective true values as EPV increased
49 (Figure S2). By design, uniform shrinkage had the same c-statistics as regular maximum
50 likelihood. When predictors were correlated, classical ridge and [Harrell's](#) PML had
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 highest c-statistics. When predictors were uncorrelated and no noise predictors were
10 present, LASSO had lower c-statistics than the maximum likelihood model. Adaptive
11 LASSO only had better discrimination than maximum likelihood when noise predictors
12
13 LASSO only had better discrimination than maximum likelihood when noise predictors
14 were present. Firth's correction did not improve the c-statistic.
15
16
17
18
19
20
21

22 3.2. Variability in the applied shrinkage 23 24 25 26 27

28 For the scenarios with five true predictors, pairwise correlations of 0.5 between
29 predictors, and an event rate of 50%, box plots of the calibration slopes over the 1 000
30 simulation runs are shown in Figure 2. For all scenarios, box plots are given in Figure
31 S3, and MAD in Figure S4. The variability of the calibration slope after shrinkage was
32 larger than the variability based on maximum likelihood, except when Firth's correction
33 was used. Firth's correction consistently reduced variability (Figure S4). This increased
34 variability was particularly strong when EPV is low, and correlations between
35 predictors were low. Only when there were 10 true predictors with high
36 intercorrelations, most shrinkage methods had lower variability than maximum
37 likelihood.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Generally, shrinkage methods improved the RMSD relative to the maximum likelihood
10 model (Table S3, Figure 3, Figure S5). However, LASSO, adaptive LASSO, classical
11 ridge and Harrell's PML often had higher RMSD than maximum likelihood when
12
13 predictors were uncorrelated and EPV or sample size was low. Classical ridge and
14
15
16
17
18 Harrell's PML often showed higher RMSD than other methods when predictors were
19 correlated and EPV was high. Two methods, the bootstrap uniform shrinkage and
20
21 Firth's correction, always had lower RMSD than maximum likelihood.
22
23
24
25
26
27
28

29 Box plots of the c-statistics also showed high between-sample variability for all
30 methods (Figure S6).
31
32
33
34
35
36

37 3.3. Correlation between estimated and optimal shrinkage

38
39
40
41
42

43 The Spearman correlation between estimated and optimal shrinkage was typically
44 negative (Figure 4, Table S4, Figures S7-8). Firth's correction was the exception with
45 consistently positive correlations. LASSO-based methods typically had the lowest
46
47 negative correlations (closest to zero). For these methods, correlations were highest,
48
49 and in particular cases even positive, in settings with more highly correlated predictors.
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 The highest positive correlations between estimated and optimal shrinkage were found
10 when there were 10 true predictors, there was non-zero true correlation between the
11 predictors and the EPV was low.
12
13
14
15
16
17
18

19 3.4. Results for coefficient estimates and variable selection 20 21 22 23 24

25 Coefficient estimates of true predictors were exaggerated when the maximum likelihood
26 model was used (Figure S9). The bias decreased with increasing EPV. Using Firth's
27 correction removed the bias. All other shrinkage methods induced negative bias and
28 consistently underestimated the coefficients. With respect to noise predictors, classical
29 ridge, Harrell's PML, LASSO, and adaptive LASSO had positive bias in the estimated
30 coefficients when there was correlation between predictors (Figure S10).
31
32
33
34
35
36
37
38
39
40
41
42

43 Regarding variable selection, adaptive LASSO selected less predictors than standard
44 LASSO implementations (Figure S11). In simulation scenarios with noise predictors,
45 these predictors were selected more often with increasing EPV, except when adaptive
46 LASSO was used (Figure S12). Table S5 summarizes how often these methods selected
47 no variables at all.
48
49
50
51
52
53
54
55
56
57
58
59
60

4. Discussion

In this paper, we assessed the performance of various shrinkage methods for clinical risk prediction models using simulations. Our key results were the following. First, shrinkage led to calibration slopes that were on average closer to the ideal value of unity than maximum likelihood. Firth's correction improved the slope least among the considered methods. Classical ridge, and to a lesser extent [Harrell's](#) PML and LASSO, tended to shrink too much overall. Second, the performance of the shrinkage methods was highly variable, especially when sample size was relatively low. The exception was Firth's correction, which showed remarkably stable performance. Despite the increased variance, the RMSD of the calibration slopes was usually lower for shrinkage methods compared to standard maximum likelihood. This was notably the case for Firth's correction, due to its limited variability, but also for bootstrap uniform shrinkage. Third, we commonly observed that the estimated shrinkage was inversely correlated with the optimal shrinkage. This corroborated the early observation by van Houwelingen,²⁰ and implies that shrinkage often does least when it is needed most. Firth's correction was again the exception, with consistently positive correlations. Fourth, there were differences between the shrinkage methods. A key parameter to this end is the RMSD,

1
2
3
4
5
6
7
8
9 because it combines bias in and variability of the calibration slope. Based on RMSD,
10
11 Firth's correction and bootstrap uniform shrinkage would be the preferred methods.
12
13 Shrinkage using the bootstrap uniform shrinkage factor performed remarkably well,
14
15 perhaps because this method explicitly uses the calibration slope for shrinkage
16
17 estimation. Firth's penalized likelihood almost surely improved performance over
18
19 maximum likelihood, with low variability and positive correlation with optimal
20
21 shrinkage. Important advantages of Firth's correction that lead to its stability are that it
22
23 does not require the estimation of a tuning parameter, and that it shrinks extreme risk
24
25 estimates. However, the magnitude of shrinkage was small.
26
27
28
29
30
31
32

33 These results have implications. Although shrinkage works on average by bringing the
34
35 calibration slope closer to unity, it may not work as anticipated for any given dataset.
36
37 The variability in the estimated shrinkage was particularly high when sample size was
38
39 low. Thus, the use of shrinkage does not justify using lower sample size for the
40
41 development of prediction models. When sample size is low, it may even be advisable
42
43 not to build a prediction model. Alternatively, a less complicated model can be
44
45 considered, for example by discarding many predictors a priori. In a previous study in
46
47 the context of survival prediction models,¹⁴ the authors suggested that it may be
48
49 possible to develop an acceptable model with EPV of 2.5 if methods like ridge or
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 LASSO are used, although acknowledged that more work was required.¹⁴ We cannot
10 defend this suggestion based on our results.
11
12
13

14
15
16
17 We have to be careful about recommendations with respect to specific shrinkage
18 approaches, because the study was not designed to inform fully on their relative merits.
19 For example, classical ridge with tuning based on 10-fold cross-validation led to poorer
20 median calibration slopes than Harrell's ~~penalized maximum likelihood~~ PML estimation
21 with tuning based on the corrected Akaike Information Criterion, but had less variability
22 in the calibration slope (Figure S4). More research should study the impact of specific
23 combinations of shrinkage and tuning methods.
24
25
26
27
28
29
30
31
32
33
34
35
36

37 The first limitation of our study was the focus on low-dimensional settings for which
38 predictors were largely pre-specified. It would be relevant to investigate the issues of
39 high variability and negative correlation in high-dimensional settings, settings where
40 both sample size and the number of potential predictors are large (such as in some
41 electronic health record studies). Second, we focused on normally distributed predictors,
42 although typical applications also contain non-normal predictors such as skewed
43 continuous predictors or categorical predictors. However, this does not invalidate the
44 key results of our paper. We anticipate the performance variability to become even be
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 larger when a mixture of predictor types are used. Third, we deliberately fixed event
10 rates in simulated datasets, because in prediction model applications one has to go by
11 the event rate that is observed in the data at hand. A downside of this choice is that our
12 results ignore sampling variability in the event rate in observational cross-sectional or
13 cohort studies. Such variability in the observed event rate may further worsen variability
14 in performance. Finally, we investigated many well-known shrinkage methods.
15
16 Nevertheless, it may be interesting to investigate whether our findings can be confirmed
17 in other approaches, such as elastic net, smoothly clipped absolute deviation (SCAD),
18 weighted fusion, or machine learning methods.³⁷⁻³⁹
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33 Our results are in line with previous work. In line with a large recent simulation study,
34 model performance in our study was related to event rate even when EPV was fixed.⁹
35 Further, the results are consistent with the recommendation to base sample size on a
36 maximal expected level of shrinkage.¹⁰ In accordance with earlier work, we observed
37 that methods like ridge or LASSO may have the tendency to shrink too much on
38 average.^{8,14-16} Perhaps the use of cross-validation may contribute to this, because
39 shrinkage parameter tuning is based on datasets with reduced sample size. However, in
40 contrast with earlier claims,^{6,14} the bootstrap uniform shrinkage method performed
41 relatively well in our simulations. These claims were based on simulations with 2.5
42 EPV, which is lower than the values considered in our study. Our results do not support
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 the development of prediction models with such low EPV with any method, although
10 more work on settings with very low event rates may be of interest.
11
12
13
14
15
16

17 In conclusion, shrinkage improves performance on average. The larger variability in
18 calibration slope with the use of shrinkage methods, and the negative correlation
19 between estimated and optimal shrinkage, suggest that shrinkage may not work well for
20 any given dataset. Firth's correction is a notable exception, with reduced variability and
21 a positive correlation between estimated and optimal shrinkage. However, the amount
22 of shrinkage it applied was modest. Overall, the use of shrinkage is not a solution to the
23 problem of low sample size or low EPV. In such cases, more fundamental changes are
24 needed, such as refraining from the development of a model, increasing sample size, or
25 reducing a priori the number of predictors if this is clinically acceptable.
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

41 **Declarations of Conflicting Interests**

42
43
44
45
46

47 The Authors declare that there is no conflict of interest.
48
49
50
51
52

53 **Funding**

54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12 The author(s) disclosed receipt of the following financial support for the research,
13
14 authorship, and/or publication of this article: This work was supported by the Research
15
16 Foundation – Flanders (FWO) [grant number G0B4716N]; and the Internal Funds KU
17
18 Leuven [grant number C24/15/037].
19
20
21
22
23
24

25 **ORCID iD**

26
27
28
29
30 Ben Van Calster 0000-0003-1613-7450
31

32
33 Maarten van Smeden 0000-0002-5529-1541
34

35
36 Bavo De Cock 0000-0002-1310-6336
37

38
39 Ewout W Steyerberg 0000-0002-7787-0122
40
41
42
43
44

45 **References**

- 46
47
48
49
50
51 1. Steyerberg EW. *Clinical prediction models* (2nd edition). New York: Springer,
52
53 20109.
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 2. Babyak MA. What you see may not be what you get: a brief, nontechnical
10 introduction to overfitting in regression-type models. *Psychosom Med* 2004; 66:
11 411-421.
12
13
- 14
15 3. Harrell FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved
16 prognostic prediction. *Stat Med* 1984; 3: 143-152.
17
18
- 19
20 4. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events
21 per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373-1379.
22
23
- 24
25 5. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with
26 logistic regression analysis: a comparison of selection and estimation methods in
27 small data sets. *Stat Med* 2000; 19: 1059-1079.
28
29
- 30
31 6. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk
32 prediction model when there are few events. *BMJ* 2015; 351: h3868.
33
34
- 35
36 7. Courvoisier DS, Combescure C, Agoritsas T, et al. Performance of logistic
37 regression modeling: beyond the number of events per variable, the role of data
38 structure. *J Clin Epidemiol* 2011; 64: 993-1000.
39
40
- 41
42 8. van Smeden M, de Groot JA, Moons KG, et al. Sample size for binary logistic
43 prediction models: Beyond events per variable criteria No rationale for 1 variable
44 per 10 events criterion for binary logistic regression analysis. *BMC Med Res*
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Methodol 2016; 16: 163.

- 1
2
3
4
5
6
7
8
9 9. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic
10 prediction models: Beyond events per variable criteria. *Stat Meth Med Res* 2019;
11 28: 2455-2474.
12
13
14
15
- 16 10. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a
17 multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat*
18 *Med* 2019; 38: 1276-1296.
19
20
21
22
- 23 11. Ogundimu EO, Altman DG and Collins GS. Adequate sample size for developing
24 prediction models is not simply related to events per variable. *J Clin Epidemiol*
25 2016; 76: 175-182.
26
27
28
29
- 30 12. Steyerberg EW, Eijkemans MJC and Habbema JDF. Application of shrinkage
31 techniques in logistic regression analysis: a case study. *Stat Neerl* 2001; 55: 76-88.
32
33
34
- 35 13. Steyerberg EW, Eijkemans MJC, Harrell FE Jr, et al. Prognostic modeling with
36 logistic regression analysis: in search of a sensible strategy in small data sets. *Med*
37 *Decis Making* 2001; 21: 45-56.
38
39
40
- 41 14. Ambler G, Seaman S and Omar RZ. An evaluation of penalised survival methods
42 for developing prognostic models with rare events. *Stat Med* 2012; 31: 1150-1161.
43
44
45
- 46 15. Pavlou M, Ambler G, Seaman SR, et al. Review and evaluation of penalised
47 regression methods for risk prediction in low-dimensional data with few events.
48 *Stat Med* 2016; 35: 1159-1177.
49
50
51
52
53
54
55
56
57
58
59
60

16. Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* 2017; 36: 2302-2317.
17. De Jong VMT, Eijkemans MJC, Van Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med* 2019; 38: 1601-1619.
18. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B* 1983; 45: 311-354.
19. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; 80: 27-38.
20. van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl* 2001; 55: 17-34.
21. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; 12: 55-67.
22. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 1996; 58: 267-288.
23. Musoro JZ, Zwinderman AH, Puhan MA, et al. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Med Res Methodol* 2014; 14: 116.
24. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study

- 1
2
3
4
5
6
7
8
9 by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; 23: 8794-
10 8801.
11
12
13 25. Van Houwelingen JC and le Cessie S. Predictive value of statistical models. *Stat*
14 *Med* 1990; 9: 1303-1325.
15
16 26. Schaefer RL, Roi LD and Wolfe RA. A ridge logistic estimator. *Commun Stat*
17 *Theory Methods* 1984; 13: 99-113.
18
19 27. Le Cessie S and van Houwelingen JC. Ridge estimators in logistic regression. *J R*
20 *Stat Soc C* 1992; 41: 191-201.
21
22 28. Harrell FE Jr. *Regression modeling strategies (2nd edition)*. New York: Springer,
23 2001~~5~~.
24
25 29. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; 101:
26 1418-1429.
27
28 30. Heinze G and Schemper M. A solution to the problem of separation in logistic
29 regression. *Stat Med* 2002; 21: 2409-2419.
30
31 31. Pencina MJ, D'Agostino RB Sr, Pencina KM, et al. Interpreting incremental value
32 of markers added to risk prediction models. *Am J Epidemiol* 2012; 176: 473-481.
33
34 32. Friedman JH, Hastie T and Tibshirani R. Regularization paths for generalized linear
35 models via coordinate descent. *J Stat Softw* 2010; 33: 1.
36
37 33. Hurvich CM and Tsai CL. Regression and time series model selection in small
38 samples. *Biometrika* 1989; 76: 297-307.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 34. Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic
10 regression: causes, consequences, and control. *Am J Epidemiol* 2018;187:864-870.
11
12
13 35. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk
14 models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; 74: 167-
15 176.
16
17
18 36. Cox DR. Two further applications of a model for binary regression. *Biometrika*
19 1958; 45: 562-565.
20
21
22 37. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R*
23 *Stat Soc B* 2005; 67: 301-320.
24
25
26 38. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its
27 oracle properties. *J Am Stat Assoc* 2001; 96: 1348-1360.
28
29
30 39. Daye ZJ and Jeng XJ. Shrinkage and model selection with correlated variables via
31 weighted fusion. *Comput Stat Data Anal* 2009; 53: 1284-1298.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Figure 1. Median calibration slopes for the scenarios with five true predictors.

10
11
12 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
13
14 shrinkage based on bootstrapping; L2, classical ridge (~~L2~~)-regression; PML, Harrell's
15
16 penalized maximum likelihood; L1, LASSO (~~L1~~)-regression; AL, adaptive LASSO; F,
17
18 logistic regression with Firth's correction.
19

20
21
22
23
24
25 Figure 2. Box plots of the calibration slope over the 1 000 simulation runs for scenarios
26
27 with five true predictors, no correlation between predictors, and 50% event rate. The
28
29 events per variable is indicated in the top left. The numbers at the bottom are the root
30
31 mean squared distances (RMSD) of the log of the calibration slopes. The length of the
32
33 whiskers is at most 1.5 times the interquartile range. Calibration slopes are winsorized
34
35 at 0.1 and 10 for visualization purposes.
36
37

38
39 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
40
41 shrinkage based on bootstrapping; L2, classical ridge (~~L2~~)-regression; PML, Harrell's
42
43 penalized maximum likelihood; L1, LASSO (~~L1~~)-regression; AL, adaptive LASSO; F,
44
45 logistic regression with Firth's correction.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Figure 3. Root mean squared distance (RMSD) of the logarithm of the calibration slope
10 over 1 000 simulation runs for scenarios with five true predictors.
11

12
13
14 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
15 shrinkage based on bootstrapping; L2, classical ridge (~~L2~~)-regression; PML, Harrell's
16 penalized maximum likelihood; L1, LASSO (~~L1~~)-regression; AL, adaptive LASSO; F,
17 logistic regression with Firth's correction.
18
19
20
21
22
23
24
25
26

27 Figure 4. Scatter plots of the slope after shrinkage versus the slope based on maximum
28 likelihood (no shrinkage) for the scenario with five true predictors, no correlation
29 between predictors, 50% event rate, and three events per variable. Each point represents
30 one of the 1 000 simulation runs. The blue line is the diagonal, where both slopes are
31 the same. The green lines show the ideal slope (unity). Red circles refer to simulation
32 runs where maximum likelihood resulted in a slope above unity.
33
34
35
36
37
38
39

40
41 LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on
42 bootstrapping; L2, classical ridge (~~L2~~)-regression; PML, Harrell's penalized maximum
43 likelihood; L1, LASSO (~~L1~~)-regression; AL, adaptive LASSO; F, logistic regression
44 with Firth's correction.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Box 1. Clinical illustration: prediction model for ovarian cancer diagnosis.

In 2005, the International Ovarian Tumor Analysis group published its first ultrasound-based risk prediction models to diagnose ovarian malignancy in patients that are selected for surgery.²⁴ The dataset of 1066 patients were randomly split in a development part of 754 (191 with a malignancy) and a validation part of 312 (75 with a malignancy). For the model, over 40 predictors were considered, totaling 52 parameters. The EPV was 3.7 (191/52). Data-driven variable selection was used in the context of standard logistic regression (no shrinkage), leading to a model with 12 predictors. Using the dataset from this study, the model had a calibration intercept of 0.007 and a calibration slope of 1.09 on the validation part. Contrary to expectation, the observed slope suggested mild underfitting: the estimated risks were too close to the overall outcome prevalence. If likelihood-based uniform shrinkage factor were used,²⁵ predictors coefficients would have been multiplied by 0.89. This implies a shrinkage of 11%, which seems little given the data-driven selection among 52 parameters in a dataset of moderate size. With this method, the calibration slope on the validation part would have been 1.22. Hence, shrinkage worsened the calibration of the model. Obviously, the small size of the validation part set implied considerable random variation. Nevertheless, this illustrates that a thorough assessment of the variability of standard logistic regression and alternatives based on shrinkage is important.

Table 1. Overview of the simulation factors in the full factorial simulation design.

Simulation factor	Factor levels
Events per variable	3, 5, 10, 20, 50
Predictors	five true predictors; 10 true predictors; five true and five noise predictors
Correlation between predictors	0, 0.5
Outcome event rate	0.1, 0.5

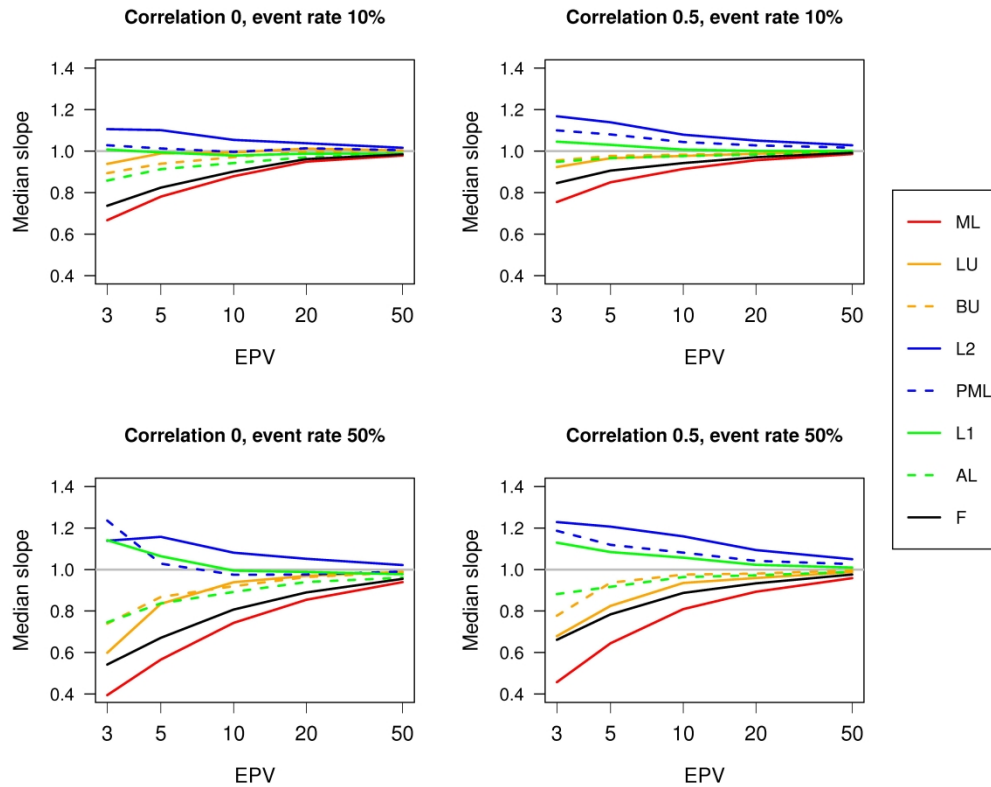
Table 2. Overview of the characteristics of the 60 simulation scenarios

Predictors	Correlation	Event rate	EPV	Events	Sample size	True c statistic	Model intercept	
Five true predictors, or five true + five noise predictors	0	0.1	3	15	150	0.75	-2.57	
			5	25	250			
			10	50	500			
			20	100	1000			
			50	250	2500			
	0.5	0.5	3	15	30	0.74	0	
			5	25	50			
			10	50	100			
			20	100	200			
			50	250	500			
	0.5	0.1	0.1	3	15	150	0.83	-2.98
				5	25	250		
				10	50	500		
				20	100	1000		
				50	250	2500		
0.5		0.5	0.5	3	15	30	0.81	0
				5	25	50		
				10	50	100		
				20	100	200		
				50	250	500		
10 true predictors	0	0.1	3	30	300	0.82	-2.88	
			5	50	500			
			10	100	1000			
			20	200	2000			
			50	500	5000			
	0.5	0.5	0.5	3	30	60	0.80	0
				5	50	100		
				10	100	200		
				20	200	400		
				50	500	1000		
	0.5	0.1	0.1	3	30	300	0.93	-4.34
				5	50	500		
				10	100	1000		
20				200	2000			

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

			50	500	5000		
		0.5	3	30	60	0.91	0
			5	50	100		
			10	100	200		
			20	200	400		
			50	500	1000		

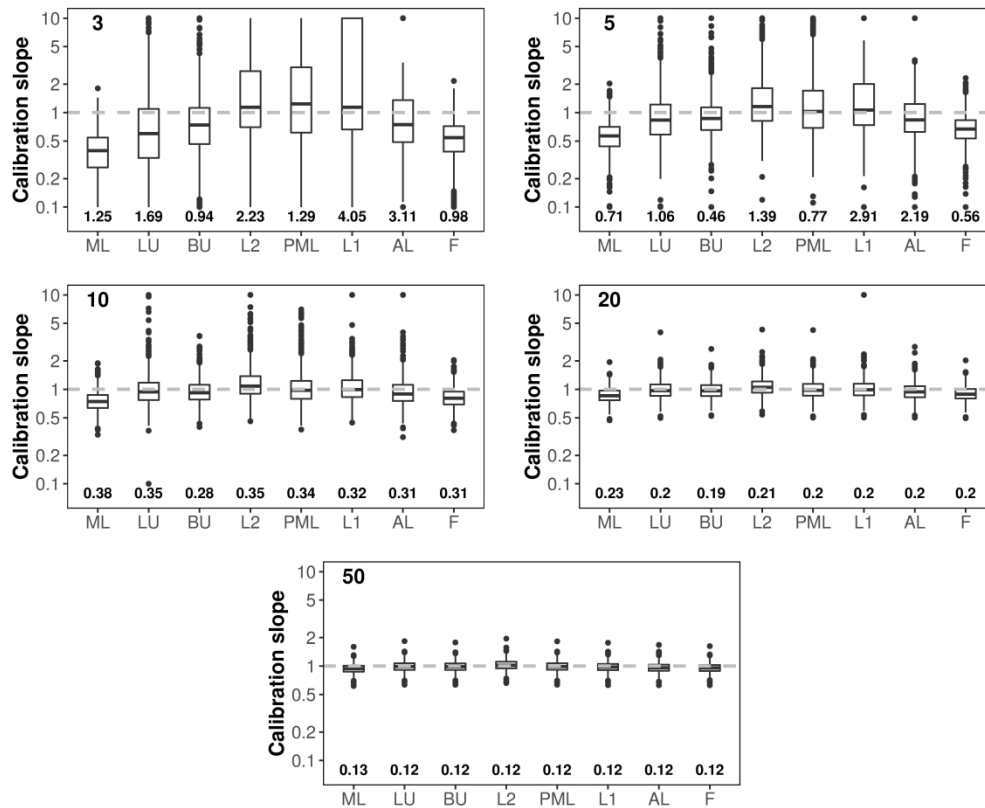
For Peer Review



Median calibration slopes for the scenarios with five true predictors.

ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrapping; L2, ridge (L2) regression; PML, penalized maximum likelihood; L1, LASSO (L1) regression; AL, adaptive LASSO; F, logistic regression with Firth's correction.

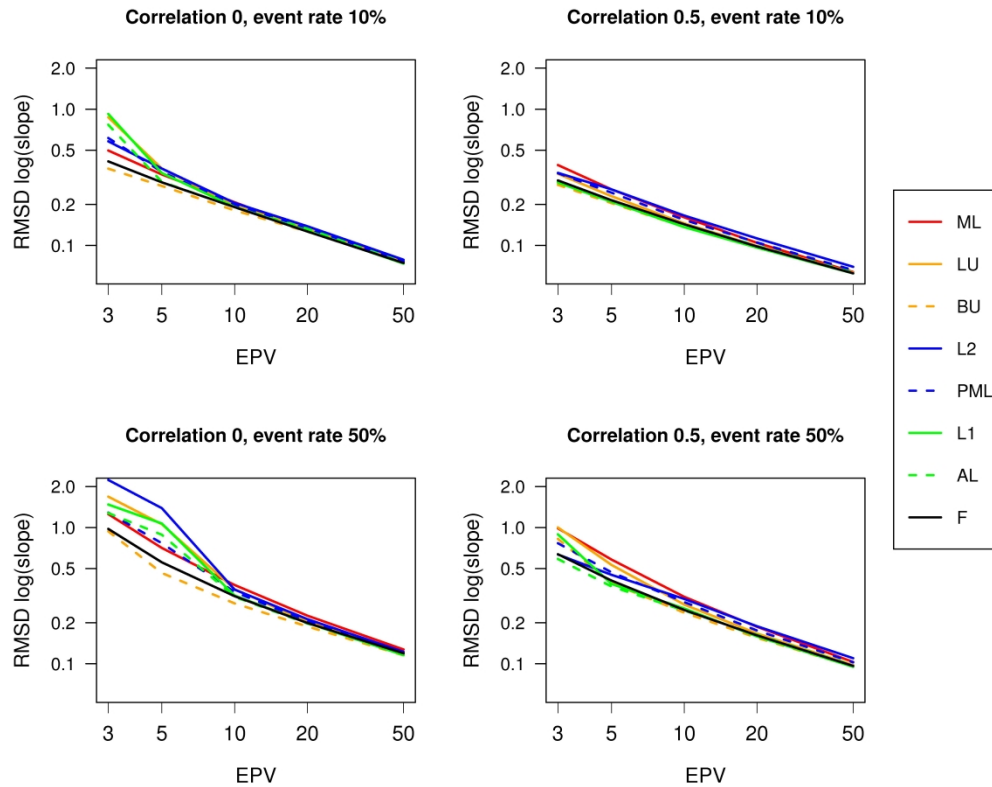
240x201mm (600 x 600 DPI)



Box plots of the calibration slope over the 1 000 simulation runs for scenarios with five true predictors, no correlation between predictors, and 50% event rate. The events per variable is indicated in the top left. The numbers at the bottom are the root mean squared distances (RMSD) of the log of the calibration slopes. The length of the whiskers is at most 1.5 times the interquartile range. Calibration slopes are winsorized at 0.1 and 10 for visualization purposes.

ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrapping; L2, ridge (L2) regression; PML, penalized maximum likelihood; L1, LASSO (L1) regression; AL, adaptive LASSO; F, logistic regression with Firth's correction.

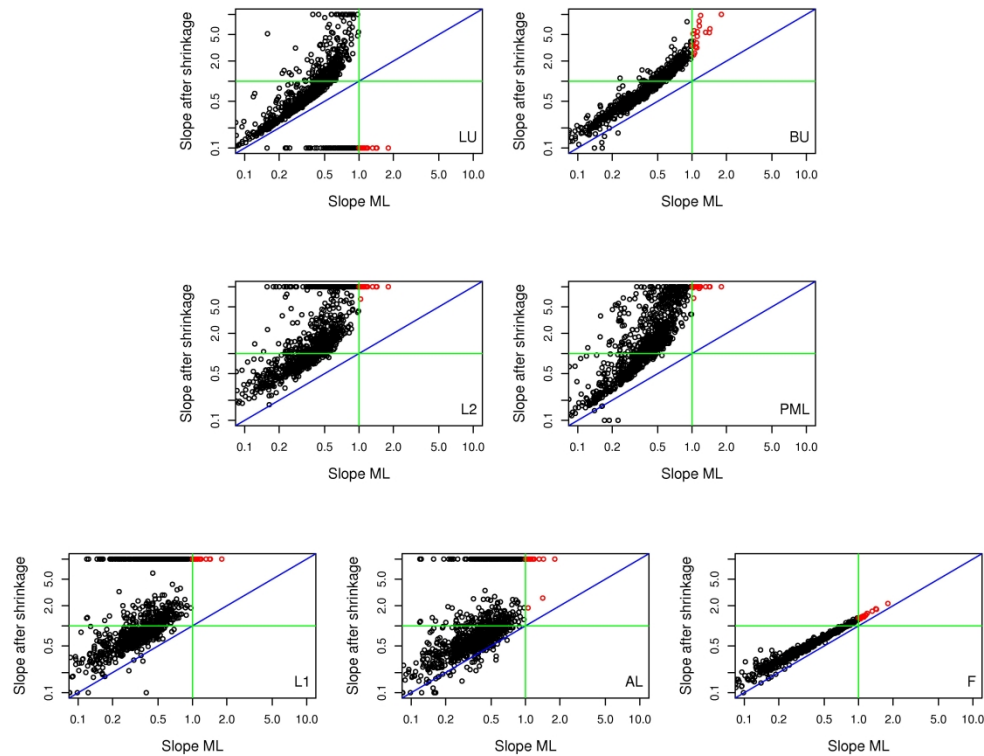
240x196mm (600 x 600 DPI)



Root mean squared distance (RMSD) of the logarithm of the calibration slope over 1 000 simulation runs for scenarios with five true predictors.

ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrapping; L2, ridge (L2) regression; PML, penalized maximum likelihood; L1, LASSO (L1) regression; AL, adaptive LASSO; F, logistic regression with Firth's correction.

240x201mm (600 x 600 DPI)



Scatter plots of the slope after shrinkage versus the slope based on maximum likelihood (no shrinkage) for the scenario with five true predictors, no correlation between predictors, 50% event rate, and three events per variable. Each point represents one of the 1 000 simulation runs. The blue line is the diagonal, where both slopes are the same. The green lines show the ideal slope (unity). Red circles refer to simulation runs where maximum likelihood resulted in a slope above unity.

LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrapping; L2, ridge (L2) regression; PML, penalized maximum likelihood; L1, LASSO (L1) regression; AL, adaptive LASSO; F, logistic regression with Firth's correction.

240x196mm (600 x 600 DPI)

1
2
3 **Regression shrinkage methods for clinical prediction models do not guarantee improved**
4
5 **performance: simulation study**
6
7

8
9
10 *Ben Van Calster, Maarten van Smeden, Bavo De Cock, Ewout W Steyerberg*
11
12

13
14 **SUPPLEMENTARY MATERIAL**
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

Table S1. Summary of simulation runs suggestive of separation, and the number of those runs that resulted in an error when fitting Harrell's PML model. Scenarios that are not mentioned did not contain runs suggestive of separation.

Simulation scenario	Affected runs, n (%)	PML error, n
3 EPV, 10 true predictors, 0.5 correlation, 0.5 event rate	124 (12%)	80
3 EPV, 5 true predictors, 0.5 correlation, 0.5 event rate	37 (4%)	25
3 EPV, 5 true predictors, 0 correlation, 0.5 event rate	8 (1%)	5
5 EPV, 10 true predictors, 0.5 correlation, 0.5 event rate	6 (1%)	2
3 EPV, 5 true + 5 noise predictors, 0.5 correlation, 0.5 event rate	2 (<1%)	1
3 EPV, 5 true + 5 noise predictors, 0 correlation, 0.5 event rate	1 (<1%)	0

Affected runs are runs where R warned that there were fitted probabilities of 0 or 1, or where the model did not converge. EPV, events per variable.

Table S2. Median calibration slope (with 5th and 95th percentile) by scenario and method.

Simulation scenario	ML	LU	BU	L2	PML	L1	AL	Firth
5 true predictors								
Corr 0, ER 10%								
EPV 3	0.67 (0.41-1.10)	0.94 (0.45-3.24)	0.89 (0.53-1.83)	1.11 (0.60-4.79)	1.03 (0.52-4.80)	1.01 (0.54-10)	0.86 (0.48-10)	0.74 (0.47-1.21)
EPV 5	0.78 (0.54-1.18)	0.99 (0.62-2.00)	0.94 (0.62-1.58)	1.10 (0.68-2.33)	1.01 (0.63-2.07)	0.99 (0.65-2.19)	0.91 (0.60-1.62)	0.82 (0.58-1.23)
EPV 10	0.88 (0.68-1.17)	0.99 (0.73-1.43)	0.97 (0.72-1.34)	1.05 (0.78-1.56)	1.00 (0.73-1.45)	0.98 (0.73-1.44)	0.94 (0.70-1.32)	0.90 (0.69-1.20)
EPV 20	0.95 (0.78-1.17)	1.01 (0.82-1.28)	1.00 (0.81-1.25)	1.04 (0.84-1.32)	1.01 (0.82-1.28)	0.99 (0.80-1.25)	0.97 (0.79-1.24)	0.96 (0.79-1.18)
EPV 50	0.98 (0.87-1.12)	1.00 (0.89-1.16)	1.00 (0.89-1.15)	1.02 (0.90-1.17)	1.00 (0.89-1.16)	0.99 (0.88-1.13)	0.99 (0.88-1.13)	0.98 (0.87-1.13)
Corr 0.5, ER 10%								
EPV 3	0.75 (0.47-1.13)	0.92 (0.54-1.68)	0.96 (0.59-1.53)	1.17 (0.70-2.01)	1.10 (0.62-1.94)	1.05 (0.65-1.77)	0.95 (0.60-1.62)	0.85 (0.55-1.24)
EPV 5	0.85 (0.62-1.19)	0.97 (0.67-1.47)	0.98 (0.71-1.41)	1.14 (0.80-1.73)	1.08 (0.75-1.63)	1.03 (0.73-1.50)	0.97 (0.69-1.38)	0.91 (0.67-1.25)
EPV 10	0.91 (0.73-1.16)	0.98 (0.77-1.27)	0.98 (0.79-1.26)	1.08 (0.84-1.43)	1.04 (0.81-1.38)	1.01 (0.80-1.30)	0.98 (0.77-1.26)	0.94 (0.76-1.19)
EPV 20	0.96 (0.81-1.13)	0.99 (0.83-1.18)	0.99 (0.84-1.18)	1.05 (0.88-1.26)	1.03 (0.86-1.23)	1.00 (0.85-1.19)	0.98 (0.83-1.18)	0.97 (0.83-1.15)
EPV 50	0.99 (0.89-1.10)	1.00 (0.90-1.11)	1.00 (0.90-1.11)	1.03 (0.92-1.15)	1.02 (0.91-1.14)	1.00 (0.90-1.12)	0.99 (0.89-1.11)	0.99 (0.89-1.10)
Corr 0, ER 50%								
EPV 3	0.39 (0.12-0.82)	0.60 (-1.99-3.56)	0.74 (0.17-2.21)	1.14 (0.31-263)	1.24 (0.18-10.0)	1.14 (0.31-10)	0.75 (0.22-10)	0.54 (0.17-1.07)
EPV 5	0.57 (0.29-1.00)	0.83 (0.29-2.90)	0.87 (0.42-1.92)	1.16 (0.51-182)	1.03 (0.40-6.31)	1.06 (0.46-10)	0.84 (0.38-10)	0.67 (0.36-1.17)
EPV 10	0.74 (0.50-1.12)	0.94 (0.57-1.80)	0.92 (0.60-1.56)	1.08 (0.67-2.22)	0.97 (0.59-1.95)	0.99 (0.63-1.96)	0.89 (0.57-1.69)	0.81 (0.54-1.21)
EPV 20	0.85 (0.65-1.17)	0.97 (0.71-1.47)	0.96 (0.72-1.38)	1.05 (0.77-1.59)	0.98 (0.71-1.47)	0.99 (0.72-1.46)	0.94 (0.70-1.35)	0.89 (0.68-1.22)
EPV 50	0.94 (0.77-1.13)	0.99 (0.81-1.21)	0.99 (0.81-1.20)	1.02 (0.84-1.27)	0.99 (0.81-1.21)	0.98 (0.81-1.20)	0.96 (0.79-1.16)	0.95 (0.79-1.15)
Corr 0.5, ER 50%								
EPV 3	0.46 (0.17-1.01)	0.68 (0.19-2.94)	0.78 (0.19-2.14)	1.23 (0.49-3.85)	1.19 (0.26-3.83)	1.13 (0.44-10)	0.88 (0.34-3.08)	0.66 (0.29-1.33)
EPV 5	0.64 (0.33-1.11)	0.82 (0.39-1.96)	0.94 (0.47-1.78)	1.21 (0.64-2.66)	1.12 (0.51-2.44)	1.08 (0.59-2.20)	0.92 (0.51-1.77)	0.78 (0.43-1.31)
EPV 10	0.81 (0.55-1.19)	0.94 (0.60-1.51)	0.98 (0.66-1.49)	1.16 (0.75-1.86)	1.08 (0.68-1.77)	1.06 (0.70-1.66)	0.96 (0.64-1.53)	0.89 (0.61-1.30)
EPV 20	0.89 (0.69-1.16)	0.96 (0.73-1.28)	0.98 (0.76-1.28)	1.09 (0.83-1.46)	1.04 (0.78-1.41)	1.02 (0.78-1.34)	0.97 (0.74-1.31)	0.93 (0.73-1.21)
EPV 50	0.96 (0.82-1.12)	0.99 (0.84-1.17)	1.00 (0.85-1.17)	1.05 (0.89-1.25)	1.03 (0.87-1.22)	1.01 (0.86-1.19)	0.99 (0.84-1.17)	0.98 (0.84-1.14)
5 true and 5 noise predictors								
Corr 0, ER 10%								
EPV 3	0.65 (0.46-0.93)	0.95 (0.57-2.02)	0.91 (0.60-1.46)	1.10 (0.67-2.27)	1.00 (0.59-2.20)	1.07 (0.64-2.47)	0.88 (0.56-1.55)	0.71 (0.50-0.99)
EPV 5	0.77 (0.59-1.02)	0.98 (0.69-1.48)	0.95 (0.71-1.35)	1.07 (0.75-1.65)	0.99 (0.70-1.54)	1.07 (0.73-1.67)	0.93 (0.68-1.37)	0.81 (0.62-1.06)
EPV 10	0.88 (0.73-1.07)	0.99 (0.79-1.27)	0.99 (0.80-1.24)	1.04 (0.84-1.34)	0.99 (0.79-1.28)	1.04 (0.82-1.40)	0.96 (0.78-1.24)	0.90 (0.74-1.09)
EPV 20	0.94 (0.83-1.09)	1.00 (0.87-1.18)	1.00 (0.87-1.17)	1.03 (0.89-1.21)	1.01 (0.87-1.18)	1.04 (0.89-1.24)	0.99 (0.86-1.16)	0.95 (0.84-1.10)
EPV 50	0.98 (0.89-1.08)	1.00 (0.91-1.11)	1.00 (0.91-1.11)	1.01 (0.92-1.12)	1.00 (0.91-1.11)	1.03 (0.93-1.16)	1.00 (0.91-1.11)	0.98 (0.90-1.08)
Corr 0.5, ER 10%								
EPV 3	0.54 (0.32-0.85)	0.79 (0.40-1.77)	0.91 (0.50-1.57)	1.14 (0.66-2.11)	1.11 (0.55-2.08)	1.05 (0.64-1.82)	0.89 (0.52-1.55)	0.67 (0.42-1.00)
EPV 5	0.70 (0.49-0.99)	0.89 (0.58-1.50)	0.95 (0.67-1.43)	1.11 (0.76-1.78)	1.08 (0.70-1.72)	1.05 (0.70-1.61)	0.93 (0.64-1.42)	0.78 (0.56-1.09)
EPV 10	0.83 (0.65-1.06)	0.94 (0.71-1.27)	0.98 (0.76-1.28)	1.07 (0.82-1.43)	1.04 (0.78-1.38)	1.04 (0.78-1.37)	0.97 (0.74-1.28)	0.88 (0.69-1.11)
EPV 20	0.91 (0.77-1.08)	0.98 (0.81-1.17)	0.99 (0.83-1.18)	1.04 (0.87-1.26)	1.02 (0.85-1.23)	1.04 (0.86-1.25)	0.99 (0.82-1.20)	0.94 (0.79-1.10)
EPV 50	0.97 (0.87-1.08)	1.00 (0.89-1.11)	1.00 (0.90-1.11)	1.02 (0.92-1.14)	1.01 (0.91-1.13)	1.03 (0.92-1.15)	1.00 (0.89-1.11)	0.98 (0.88-1.08)
Corr 0, ER 50%								

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

EPV 3	0.38 (0.19-0.66)	0.68 (0.21-2.92)	0.81 (0.39-1.66)	1.13 (0.49-1.67)	1.20 (0.37-4.35)	1.12 (0.45-1.10)	0.76 (0.32-1.10)	0.50 (0.26-0.82)
EPV 5	0.55 (0.35-0.83)	0.84 (0.46-1.90)	0.88 (0.54-1.49)	1.10 (0.63-3.24)	1.01 (0.52-2.75)	1.09 (0.60-3.03)	0.86 (0.50-1.63)	0.63 (0.42-0.94)
EPV 10	0.73 (0.56-0.98)	0.93 (0.66-1.48)	0.94 (0.70-1.38)	1.06 (0.75-1.76)	0.95 (0.67-1.58)	1.07 (0.74-1.68)	0.92 (0.66-1.39)	0.79 (0.61-1.05)
EPV 20	0.85 (0.70-1.05)	0.97 (0.77-1.26)	0.97 (0.78-1.25)	1.04 (0.82-1.36)	0.97 (0.77-1.27)	1.03 (0.81-1.39)	0.96 (0.77-1.24)	0.88 (0.72-1.09)
EPV 50	0.93 (0.82-1.07)	0.98 (0.85-1.14)	0.99 (0.86-1.15)	1.01 (0.88-1.18)	0.98 (0.85-1.14)	1.03 (0.88-1.23)	0.98 (0.84-1.14)	0.95 (0.83-1.08)
Corr 0.5, ER 50%								
EPV 3	0.45 (0.23-0.76)	0.67 (0.29-1.59)	0.91 (0.42-1.67)	1.17 (0.62-2.26)	1.15 (0.49-2.28)	1.07 (0.58-2.17)	0.87 (0.47-1.65)	0.61 (0.35-0.97)
EPV 5	0.64 (0.42-0.95)	0.82 (0.49-1.46)	0.96 (0.63-1.53)	1.13 (0.72-1.90)	1.09 (0.65-1.84)	1.09 (0.69-1.74)	0.94 (0.59-1.52)	0.74 (0.50-1.10)
EPV 10	0.80 (0.62-1.04)	0.92 (0.69-1.26)	0.99 (0.75-1.30)	1.09 (0.82-1.47)	1.05 (0.78-1.42)	1.06 (0.77-1.42)	0.97 (0.71-1.30)	0.87 (0.67-1.10)
EPV 20	0.90 (0.75-1.08)	0.96 (0.80-1.18)	0.99 (0.83-1.21)	1.05 (0.87-1.29)	1.02 (0.84-1.26)	1.05 (0.86-1.28)	0.99 (0.81-1.21)	0.93 (0.78-1.11)
EPV 50	0.96 (0.85-1.07)	0.99 (0.87-1.10)	1.00 (0.89-1.11)	1.03 (0.91-1.15)	1.01 (0.89-1.13)	1.03 (0.91-1.17)	0.99 (0.88-1.12)	0.97 (0.86-1.08)
10 true predictors								
Corr 0, ER 10%								
EPV 3	0.73 (0.54-0.98)	0.91 (0.62-1.41)	0.96 (0.70-1.34)	1.06 (0.75-1.62)	0.94 (0.63-1.48)	0.97 (0.70-1.47)	0.88 (0.63-1.26)	0.80 (0.59-1.06)
EPV 5	0.83 (0.66-1.04)	0.95 (0.73-1.29)	0.98 (0.77-1.27)	1.05 (0.81-1.39)	0.96 (0.73-1.31)	0.98 (0.75-1.28)	0.92 (0.72-1.22)	0.87 (0.70-1.08)
EPV 10	0.91 (0.77-1.07)	0.98 (0.81-1.17)	0.99 (0.83-1.17)	1.03 (0.86-1.23)	0.98 (0.81-1.17)	0.97 (0.82-1.17)	0.95 (0.80-1.12)	0.93 (0.79-1.09)
EPV 20	0.95 (0.85-1.07)	0.99 (0.88-1.12)	0.99 (0.89-1.12)	1.01 (0.90-1.14)	0.99 (0.88-1.12)	0.98 (0.87-1.10)	0.96 (0.86-1.09)	0.96 (0.86-1.08)
EPV 50	0.98 (0.91-1.06)	1.00 (0.93-1.07)	1.00 (0.93-1.08)	1.01 (0.94-1.09)	1.00 (0.92-1.07)	0.99 (0.92-1.07)	0.99 (0.92-1.06)	0.99 (0.92-1.06)
Corr 0.5, ER 10%								
EPV 3	0.77 (0.52-1.05)	0.85 (0.57-1.21)	1.01 (0.68-1.36)	1.16 (0.83-1.58)	1.08 (0.70-1.50)	1.06 (0.77-1.39)	0.96 (0.68-1.28)	0.89 (0.63-1.17)
EPV 5	0.86 (0.68-1.08)	0.91 (0.71-1.16)	1.01 (0.79-1.25)	1.12 (0.89-1.40)	1.05 (0.81-1.34)	1.04 (0.83-1.30)	0.99 (0.76-1.23)	0.93 (0.74-1.15)
EPV 10	0.93 (0.79-1.07)	0.96 (0.81-1.12)	1.01 (0.85-1.16)	1.08 (0.92-1.25)	1.04 (0.86-1.21)	1.02 (0.87-1.18)	1.00 (0.85-1.16)	0.97 (0.82-1.11)
EPV 20	0.97 (0.86-1.08)	0.98 (0.87-1.10)	1.01 (0.90-1.12)	1.05 (0.93-1.17)	1.02 (0.91-1.15)	1.01 (0.90-1.14)	1.00 (0.88-1.12)	0.99 (0.88-1.10)
EPV 50	0.99 (0.92-1.06)	0.99 (0.93-1.07)	1.00 (0.94-1.07)	1.02 (0.96-1.09)	1.01 (0.94-1.08)	1.00 (0.94-1.07)	0.99 (0.93-1.07)	0.99 (0.93-1.07)
Corr 0, ER 50%								
EPV 3	0.44 (0.24-0.75)	0.67 (0.32-1.77)	0.90 (0.45-1.66)	1.09 (0.58-3.13)	0.96 (0.37-3.34)	0.94 (0.51-1.10)	0.73 (0.40-1.54)	0.59 (0.36-0.95)
EPV 5	0.62 (0.41-0.91)	0.82 (0.50-1.51)	0.94 (0.63-1.49)	1.08 (0.68-1.94)	0.90 (0.52-1.83)	0.96 (0.63-1.85)	0.83 (0.55-1.39)	0.73 (0.05-1.05)
EPV 10	0.79 (0.62-1.03)	0.92 (0.69-1.29)	0.97 (0.75-1.32)	1.05 (0.79-1.47)	0.93 (0.69-1.32)	0.97 (0.75-1.39)	0.91 (0.69-1.25)	0.85 (0.67-1.11)
EPV 20	0.89 (0.74-1.07)	0.96 (0.78-1.18)	0.99 (0.81-1.20)	1.04 (0.85-1.29)	0.96 (0.79-1.18)	0.98 (0.81-1.20)	0.94 (0.78-1.16)	0.92 (0.77-1.10)
EPV 50	0.95 (0.85-1.08)	0.98 (0.87-1.12)	0.99 (0.88-1.13)	1.01 (0.90-1.15)	0.98 (0.87-1.12)	0.98 (0.87-1.10)	0.96 (0.86-1.09)	0.97 (0.86-1.09)
Corr 0.5, ER 50%								
EPV 3	0.46 (0.23-0.83)	0.57 (0.27-1.16)	0.78 (0.28-1.63)	1.22 (0.64-1.95)	1.13 (0.31-1.90)	1.08 (0.60-1.73)	0.87 (0.45-1.41)	0.72 (0.37-1.14)
EPV 5	0.66 (0.38-0.97)	0.76 (0.42-1.17)	1.03 (0.57-1.46)	1.19 (0.78-1.67)	1.07 (0.59-1.61)	1.07 (0.71-1.49)	0.92 (0.61-1.33)	0.83 (0.54-1.15)
EPV 10	0.82 (0.62-1.06)	0.89 (0.66-1.16)	1.01 (0.77-1.27)	1.14 (0.86-1.45)	1.05 (0.77-1.39)	1.05 (0.82-1.35)	0.97 (0.74-1.25)	0.91 (0.70-1.16)
EPV 20	0.91 (0.74-1.09)	0.94 (0.77-1.14)	1.00 (0.82-1.20)	1.08 (0.89-1.29)	1.03 (0.84-1.25)	1.03 (0.85-1.23)	0.99 (0.82-1.18)	0.96 (0.79-1.14)
EPV 50	0.97 (0.86-1.06)	0.98 (0.87-1.08)	1.00 (0.89-1.11)	1.04 (0.92-1.16)	1.02 (0.90-1.12)	1.01 (0.90-1.12)	0.99 (0.88-1.10)	0.98 (0.87-1.08)

Corr, correlation; ER, event rate; EPV, events per variable; ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO.

Table S3. Root mean squared distance (RMSD) of the logarithm of the calibration slope by scenario and method.

Simulation scenario	ML	LU	BU	L2	PML	L1	AL	Firth
5 true predictors								
Corr 0, ER 10%								
EPV 3	0.50	0.87	0.37	0.58	0.61	0.92	0.77	0.41
EPV 5	0.33	0.37	0.27	0.37	0.35	0.34	0.29	0.29
EPV 10	0.21	0.20	0.18	0.21	0.20	0.19	0.19	0.19
EPV 20	0.13	0.13	0.13	0.14	0.13	0.13	0.13	0.13
EPV 50	0.08	0.08	0.08	0.08	0.08	0.07	0.07	0.07
Corr 0.5, ER 10%								
EPV 3	0.39	0.34	0.28	0.34	0.34	0.29	0.29	0.30
EPV 5	0.26	0.23	0.20	0.26	0.24	0.21	0.21	0.21
EPV 10	0.16	0.15	0.14	0.17	0.15	0.14	0.14	0.14
EPV 20	0.10	0.10	0.10	0.11	0.11	0.10	0.10	0.10
EPV 50	0.06	0.06	0.06	0.07	0.07	0.06	0.06	0.06
Corr 0, ER 50%								
EPV 3	1.25	1.69	0.94	2.23	1.29	1.47	1.28	0.98
EPV 5	0.71	1.06	0.46	1.39	0.77	1.07	0.88	0.56
EPV 10	0.38	0.35	0.28	0.35	0.34	0.32	0.31	0.31
EPV 20	0.23	0.20	0.19	0.21	0.20	0.20	0.20	0.20
EPV 50	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Corr 0.5, ER 50%								
EPV 3	0.99	1.00	0.83	0.63	0.77	0.89	0.59	0.64
EPV 5	0.58	0.53	0.39	0.45	0.47	0.39	0.37	0.41
EPV 10	0.31	0.27	0.24	0.30	0.28	0.25	0.25	0.25
EPV 20	0.19	0.17	0.16	0.19	0.18	0.16	0.16	0.16
EPV 50	0.10	0.10	0.10	0.11	0.10	0.09	0.10	0.10
5 true and 5 noise predictors								
Corr 0, ER 10%								
EPV 3	0.48	0.38	0.27	0.37	0.37	0.38	0.31	0.41
EPV 5	0.30	0.22	0.19	0.24	0.23	0.24	0.21	0.26
EPV 10	0.17	0.14	0.13	0.14	0.14	0.16	0.14	0.16
EPV 20	0.10	0.09	0.09	0.10	0.09	0.11	0.09	0.09
EPV 50	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.06
Corr 0.5, ER 10%								
EPV 3	0.69	0.50	0.35	0.36	0.38	0.31	0.34	0.48
EPV 5	0.42	0.29	0.22	0.26	0.27	0.24	0.24	0.32
EPV 10	0.23	0.17	0.15	0.17	0.17	0.16	0.16	0.19
EPV 20	0.13	0.11	0.10	0.11	0.11	0.11	0.11	0.12
EPV 50	0.07	0.06	0.06	0.07	0.06	0.07	0.06	0.06
Corr 0, ER 50%								
EPV 3	1.09	1.08	0.52	1.49	0.77	1.09	0.90	0.84
EPV 5	0.65	0.55	0.32	0.47	0.48	0.48	0.39	0.52
EPV 10	0.35	0.24	0.20	0.24	0.24	0.25	0.22	0.29
EPV 20	0.20	0.15	0.13	0.15	0.15	0.16	0.15	0.17
EPV 50	0.11	0.09	0.09	0.09	0.09	0.10	0.09	0.10
Corr 0.5, ER 50%								
EPV 3	0.91	0.67	0.44	0.40	0.47	0.37	0.39	0.60
EPV 5	0.52	0.37	0.26	0.31	0.32	0.28	0.29	0.38
EPV 10	0.27	0.20	0.16	0.19	0.19	0.18	0.18	0.21
EPV 20	0.15	0.12	0.11	0.13	0.12	0.13	0.12	0.13
EPV 50	0.08	0.07	0.07	0.07	0.07	0.08	0.07	0.07
10 true predictors								
Corr 0, ER 10%								
EPV 3	0.36	0.24	0.20	0.23	0.24	0.21	0.24	0.28
EPV 5	0.23	0.17	0.15	0.16	0.17	0.16	0.17	0.19
EPV 10	0.14	0.11	0.10	0.11	0.11	0.11	0.11	0.12
EPV 20	0.08	0.07	0.07	0.07	0.07	0.07	0.08	0.08
EPV 50	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Corr 0.5, ER 10%								
EPV 3	0.34	0.28	0.20	0.23	0.23	0.18	0.19	0.22
EPV 5	0.21	0.18	0.14	0.18	0.16	0.14	0.14	0.15
EPV 10	0.12	0.10	0.09	0.12	0.10	0.09	0.09	0.10
EPV 20	0.08	0.07	0.07	0.08	0.07	0.07	0.07	0.07
EPV 50	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04

Corr 0, ER 50%								
EPV 3	0.90	0.65	0.41	0.49	0.63	0.73	0.49	0.61
EPV 5	0.54	0.37	0.26	0.32	0.37	0.30	0.32	0.39
EPV 10	0.28	0.20	0.17	0.19	0.20	0.18	0.19	0.22
EPV 20	0.16	0.13	0.12	0.13	0.13	0.12	0.13	0.14
EPV 50	0.09	0.07	0.07	0.07	0.07	0.07	0.08	0.08
Corr 0.5, ER 50%								
EPV 3	0.90	0.72	0.62	0.37	0.55	0.32	0.42	0.51
EPV 5	0.53	0.43	0.29	0.28	0.31	0.23	0.26	0.31
EPV 10	0.26	0.21	0.15	0.19	0.18	0.15	0.16	0.18
EPV 20	0.15	0.13	0.11	0.13	0.12	0.11	0.11	0.12
EPV 50	0.08	0.07	0.06	0.08	0.07	0.06	0.07	0.07

Corr, correlation; ER, event rate; EPV, events per variable; ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression ; AL, adaptive LASSO.

Table S4. Spearman correlation between estimated and optimal shrinkage by scenario and method. Correlations lower than 0 are marked in red.

Simulation scenario	LU	BU	L2	PML	L1	AL	Firth
Corr 0, ER 10%							
EPV 3	-0.71	-0.64	-0.58	-0.83	-0.42	-0.29	0.64
EPV 5	-0.90	-0.75	-0.68	-0.91	-0.39	-0.22	0.68
EPV 10	-0.94	-0.72	-0.58	-0.96	-0.17	-0.15	0.76
EPV 20	-0.97	-0.56	-0.46	-0.98	-0.05	-0.16	0.79
EPV 50	-0.97	-0.31	-0.31	-0.99	0.06	0.02	0.80
Corr 0.5, ER 10%							
EPV 3	-0.89	-0.33	-0.26	-0.74	-0.03	0.00	0.80
EPV 5	-0.92	-0.33	-0.36	-0.82	-0.01	-0.02	0.83
EPV 10	-0.95	-0.33	-0.37	-0.80	0.01	-0.06	0.85
EPV 20	-0.95	-0.19	-0.30	-0.71	0.01	-0.02	0.87
EPV 50	-0.97	-0.12	-0.23	-0.65	-0.05	0.00	0.88
Corr 0, ER 50%							
EPV 3	-0.36	-0.65	-0.43	-0.62	-0.15	-0.09	0.56
EPV 5	-0.60	-0.62	-0.54	-0.74	-0.36	-0.19	0.62
EPV 10	-0.85	-0.71	-0.58	-0.86	-0.34	-0.21	0.58
EPV 20	-0.93	-0.73	-0.52	-0.95	-0.23	-0.15	0.51
EPV 50	-0.96	-0.49	-0.43	-0.98	0.00	-0.04	0.50
Corr 0.5, ER 50%							
EPV 3	-0.75	-0.71	0.15	-0.33	0.07	0.22	0.64
EPV 5	-0.88	-0.41	-0.11	-0.59	0.00	0.15	0.80
EPV 10	-0.93	-0.42	-0.37	-0.79	-0.10	0.02	0.81
EPV 20	-0.96	-0.35	-0.38	-0.82	-0.03	-0.01	0.82
EPV 50	-0.97	-0.18	-0.25	-0.74	0.05	0.02	0.81
Corr 0, ER 10%							
EPV 3	-0.81	-0.71	-0.65	-0.83	-0.44	-0.23	0.68
EPV 5	-0.90	-0.80	-0.68	-0.91	-0.41	-0.22	0.71
EPV 10	-0.94	-0.78	-0.61	-0.95	-0.40	-0.28	0.76
EPV 20	-0.96	-0.66	-0.53	-0.97	-0.25	-0.13	0.83
EPV 50	-0.98	-0.41	-0.37	-0.98	-0.18	-0.09	0.84
Corr 0.5, ER 10%							
EPV 3	-0.83	-0.44	0.07	-0.28	0.23	0.18	0.81
EPV 5	-0.88	-0.32	-0.16	-0.53	-0.01	0.01	0.82
EPV 10	-0.92	-0.49	-0.37	-0.74	-0.09	-0.08	0.85
EPV 20	-0.95	-0.50	-0.37	-0.80	-0.12	-0.11	0.87
EPV 50	-0.96	-0.26	-0.21	-0.83	-0.05	-0.06	0.87
Corr 0, ER 50%							
EPV 3	-0.54	-0.52	-0.34	-0.52	-0.13	-0.02	0.64
EPV 5	-0.74	-0.67	-0.55	-0.74	-0.32	-0.13	0.63
EPV 10	-0.88	-0.78	-0.61	-0.90	-0.44	-0.16	0.66
EPV 20	-0.93	-0.80	-0.57	-0.95	-0.34	-0.16	0.63
EPV 50	-0.97	-0.59	-0.44	-0.98	-0.26	-0.15	0.65
Corr 0.5, ER 50%							
EPV 3	-0.81	-0.50	0.25	-0.04	0.28	0.33	0.79
EPV 5	-0.88	-0.32	-0.10	-0.39	0.08	0.08	0.82
EPV 10	-0.93	-0.52	-0.30	-0.68	-0.02	0.00	0.83
EPV 20	-0.95	-0.49	-0.33	-0.79	-0.02	-0.06	0.84
EPV 50	-0.97	-0.26	-0.23	-0.80	-0.01	-0.07	0.84
Corr 0, ER 10%							
EPV 3	-0.86	-0.49	-0.43	-0.87	-0.21	-0.06	0.79
EPV 5	-0.91	-0.58	-0.47	-0.93	-0.15	-0.11	0.84
EPV 10	-0.94	-0.52	-0.38	-0.96	0.02	-0.03	0.86
EPV 20	-0.94	-0.37	-0.26	-0.97	0.04	0.00	0.84
EPV 50	-0.95	-0.16	-0.17	-0.97	0.09	0.07	0.84
Corr 0.5, ER 10%							
EPV 3	-0.91	0.18	0.43	-0.30	0.53	0.43	0.91
EPV 5	-0.92	0.43	0.18	-0.52	0.34	0.19	0.91
EPV 10	-0.92	0.27	0.08	-0.56	0.25	0.14	0.90
EPV 20	-0.93	0.11	-0.07	-0.49	0.14	0.01	0.92
EPV 50	-0.93	0.04	-0.03	-0.42	0.14	-0.01	0.91
Corr 0, ER 50%							
EPV 3	-0.79	-0.55	-0.28	-0.73	-0.08	0.13	0.75
EPV 5	-0.85	-0.38	-0.42	-0.85	-0.15	0.02	0.78
EPV 10	-0.89	-0.57	-0.42	-0.91	-0.19	-0.09	0.78

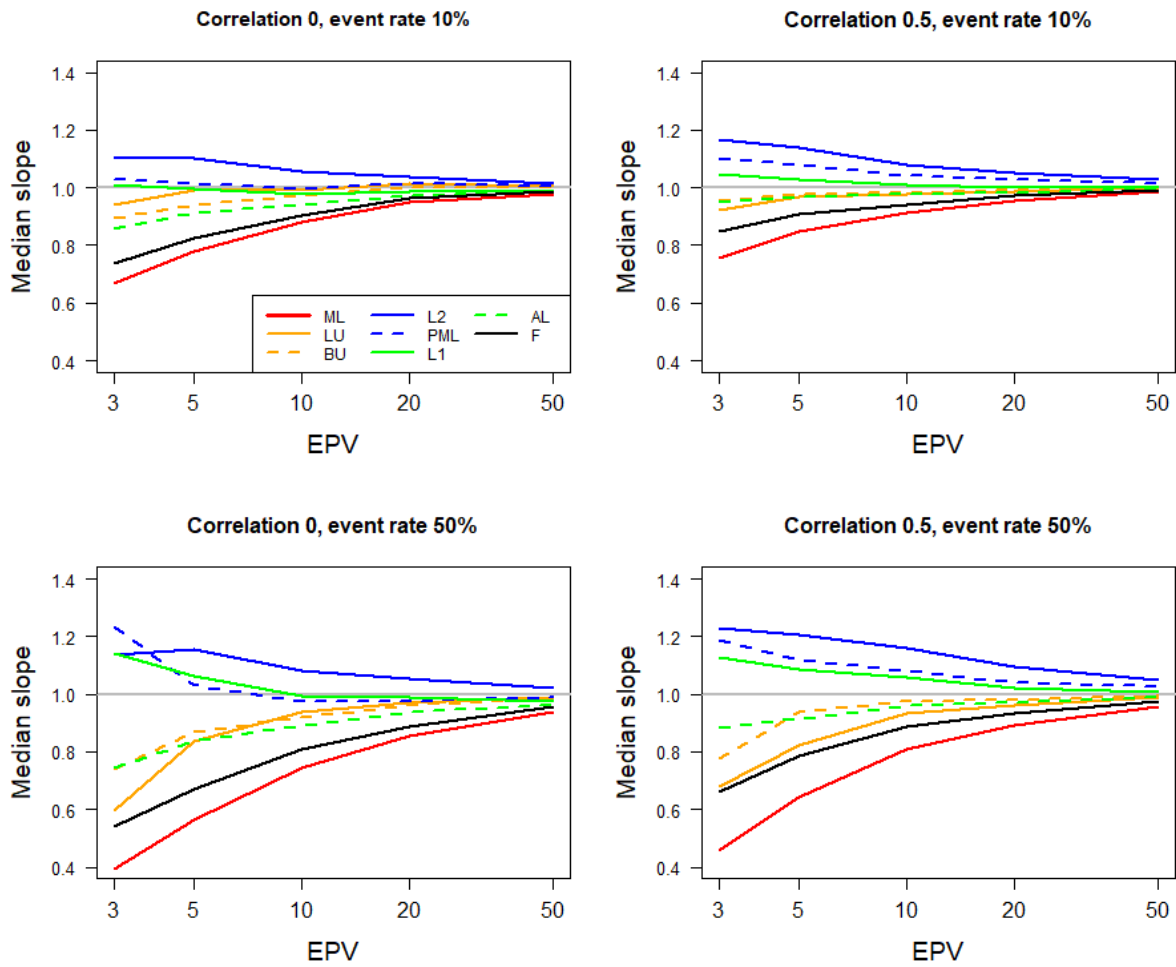
EPV 20	-0.93	-0.53	-0.34	-0.96	-0.08	-0.11	0.81
EPV 50	-0.94	-0.34	-0.23	-0.98	0.10	0.05	0.81
Corr 0.5, ER 50%							
EPV 3	-0.78	-0.56	0.59	0.25	0.58	0.54	0.61
EPV 5	-0.90	0.17	0.52	-0.13	0.57	0.44	0.86
EPV 10	-0.92	0.47	0.16	-0.53	0.34	0.25	0.92
EPV 20	-0.94	0.22	0.05	-0.59	0.26	0.17	0.92
EPV 50	-0.95	0.14	-0.01	-0.55	0.14	0.06	0.92

Corr, correlation; ER, event rate; EPV, events per variable; ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression ; AL, adaptive LASSO.

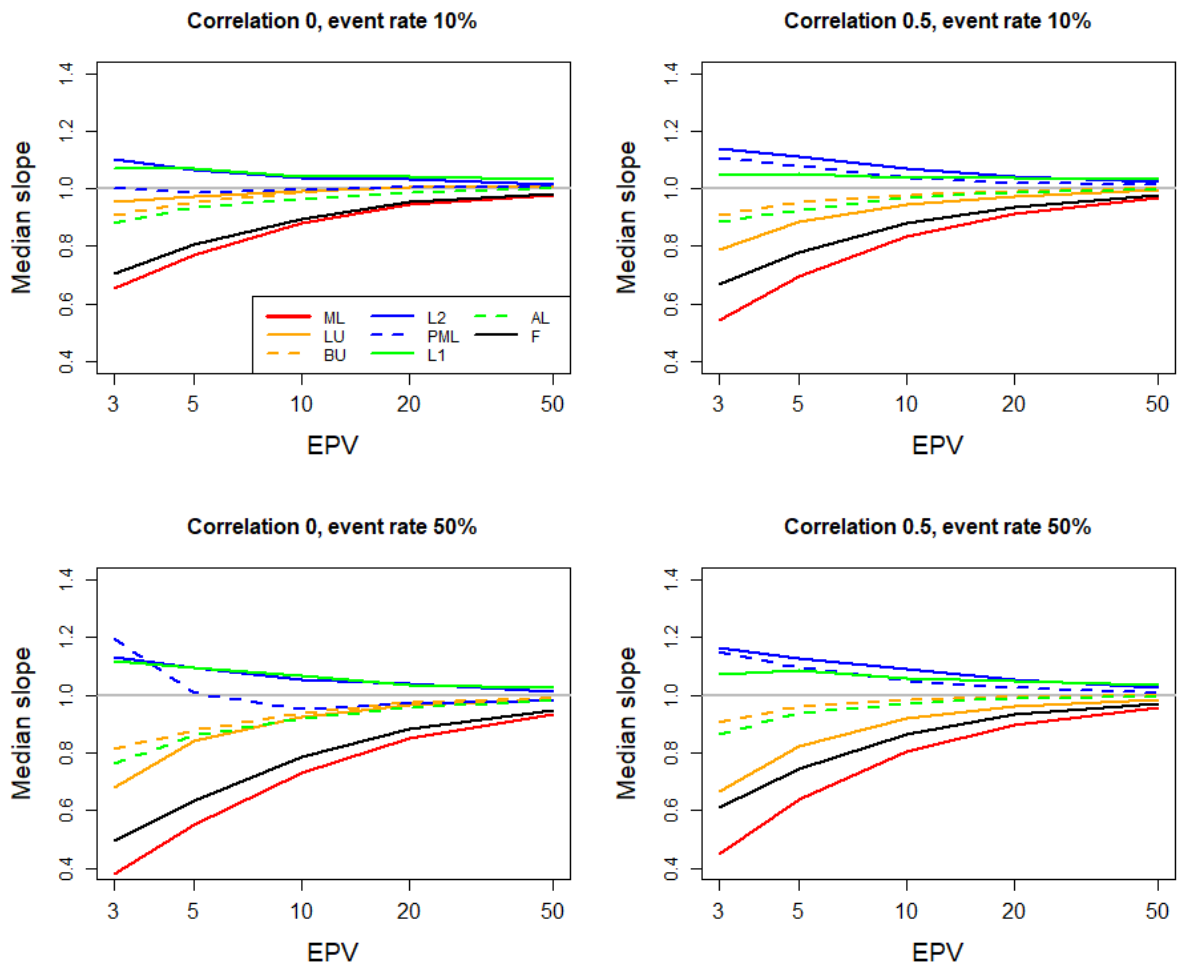
For Peer Review

Figure S1. Median calibration slopes per scenario and method. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

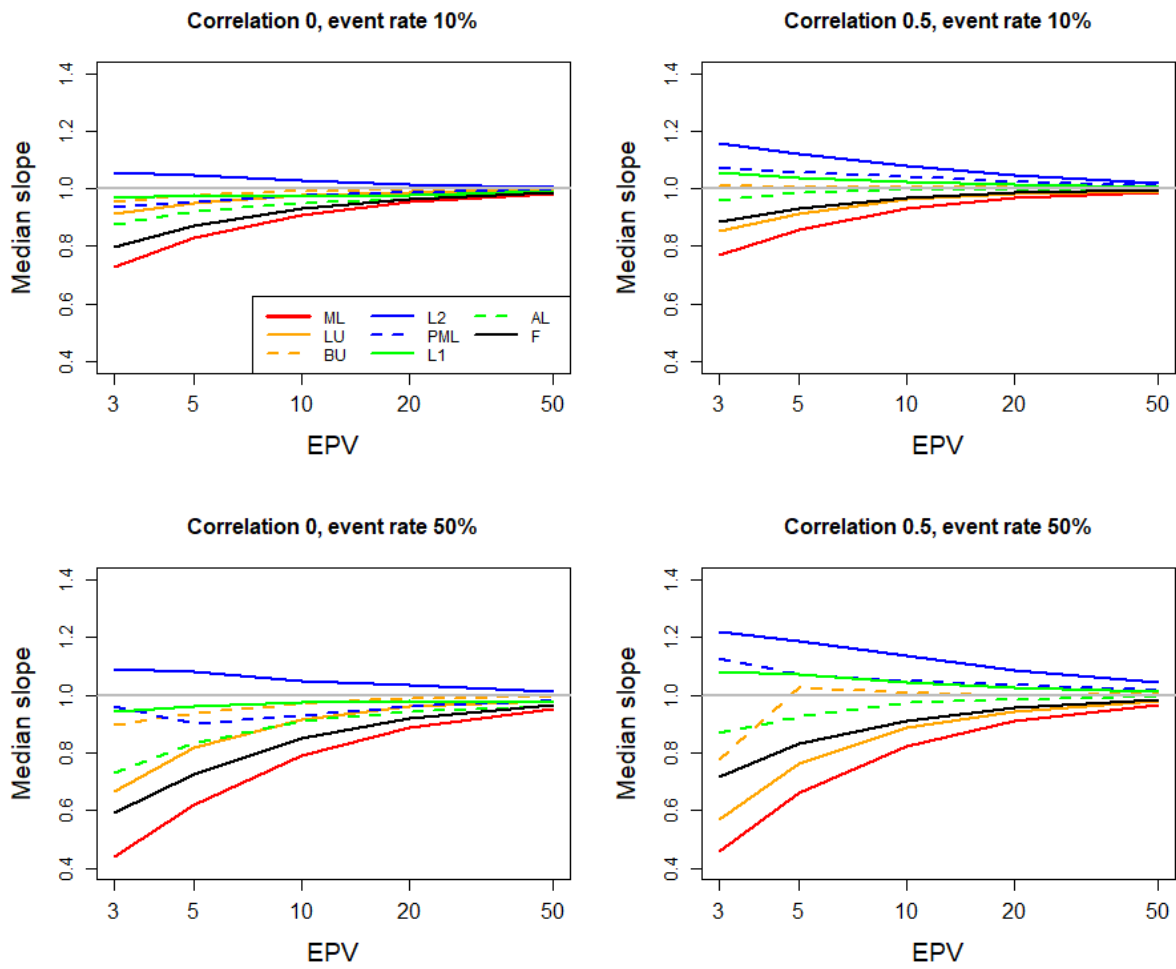


B. Scenarios with 5 true and 5 noise predictors



new

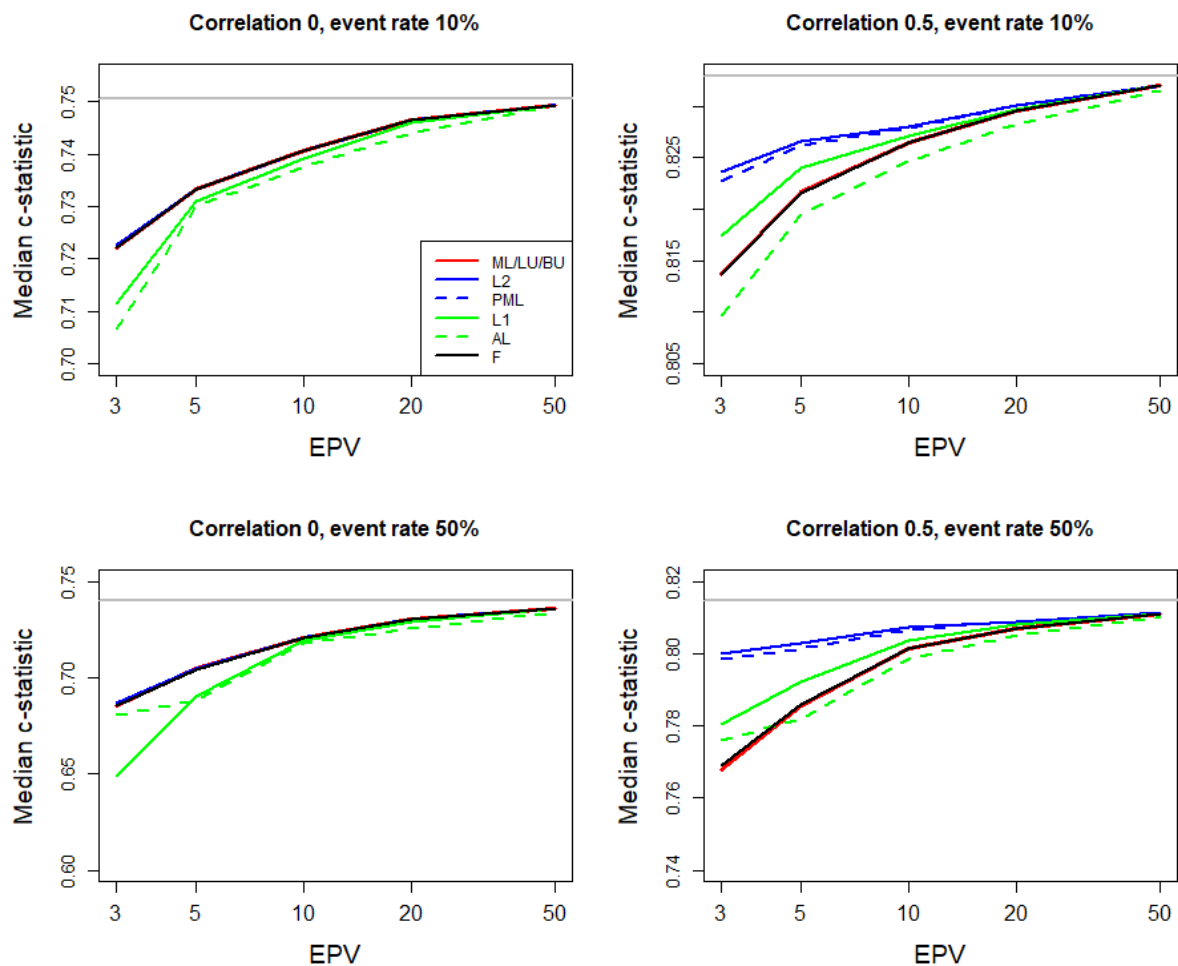
C. Scenarios with 10 true predictors



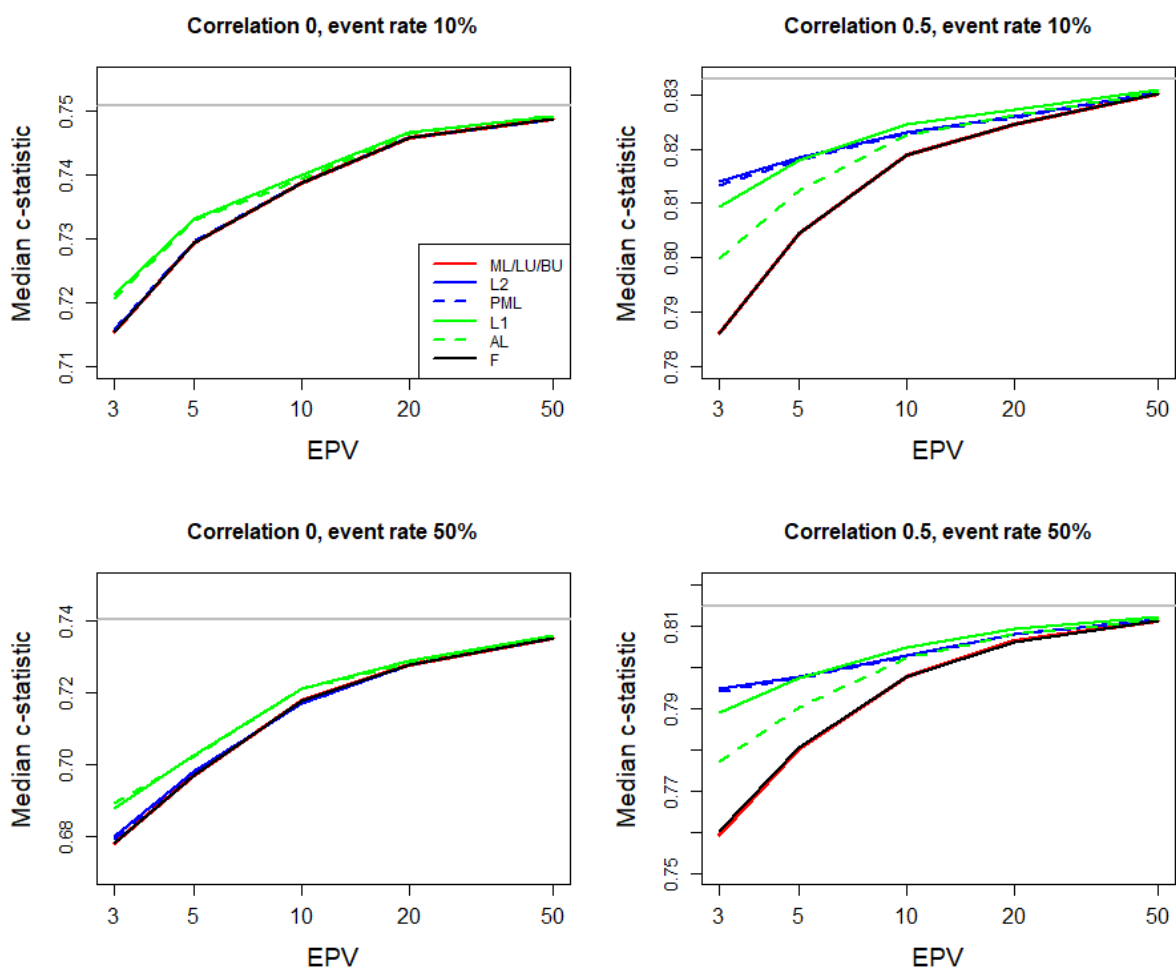
ew

Figure S2. Median c-statistic per scenario and method. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

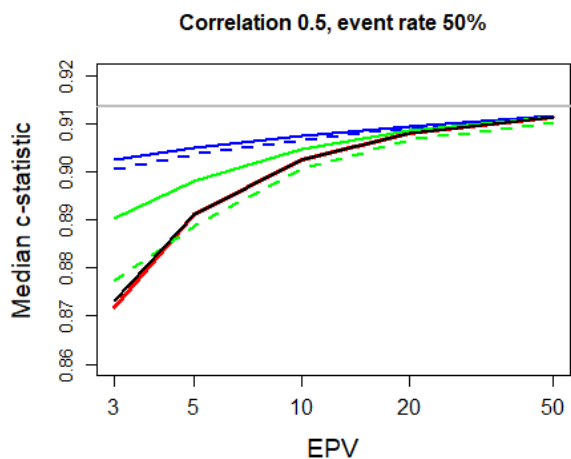
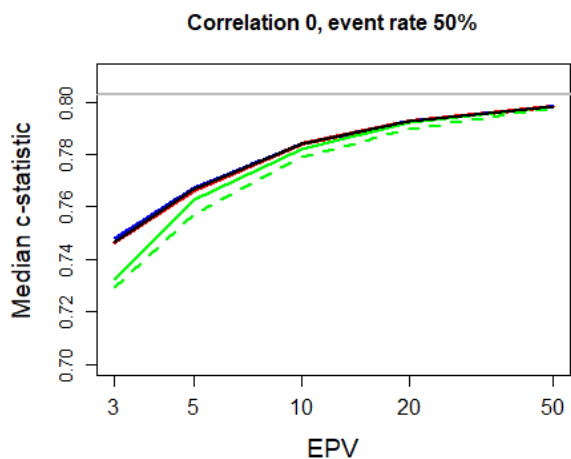
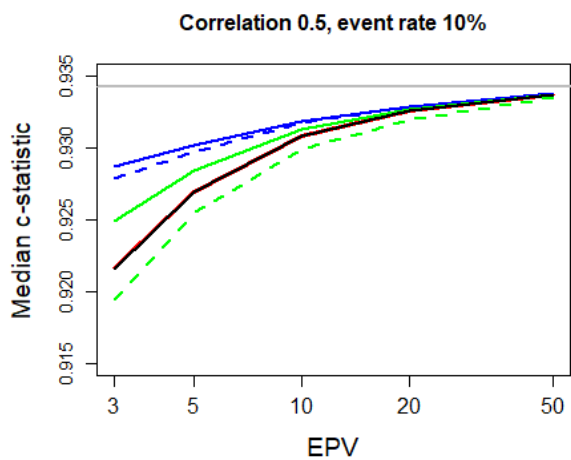
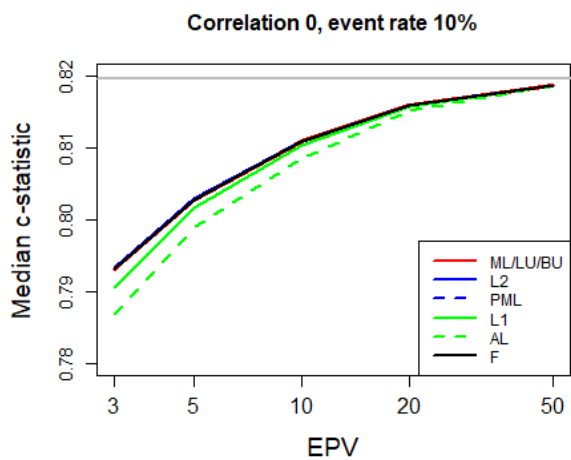


B. Scenarios with 5 true and 5 noise predictors



ew

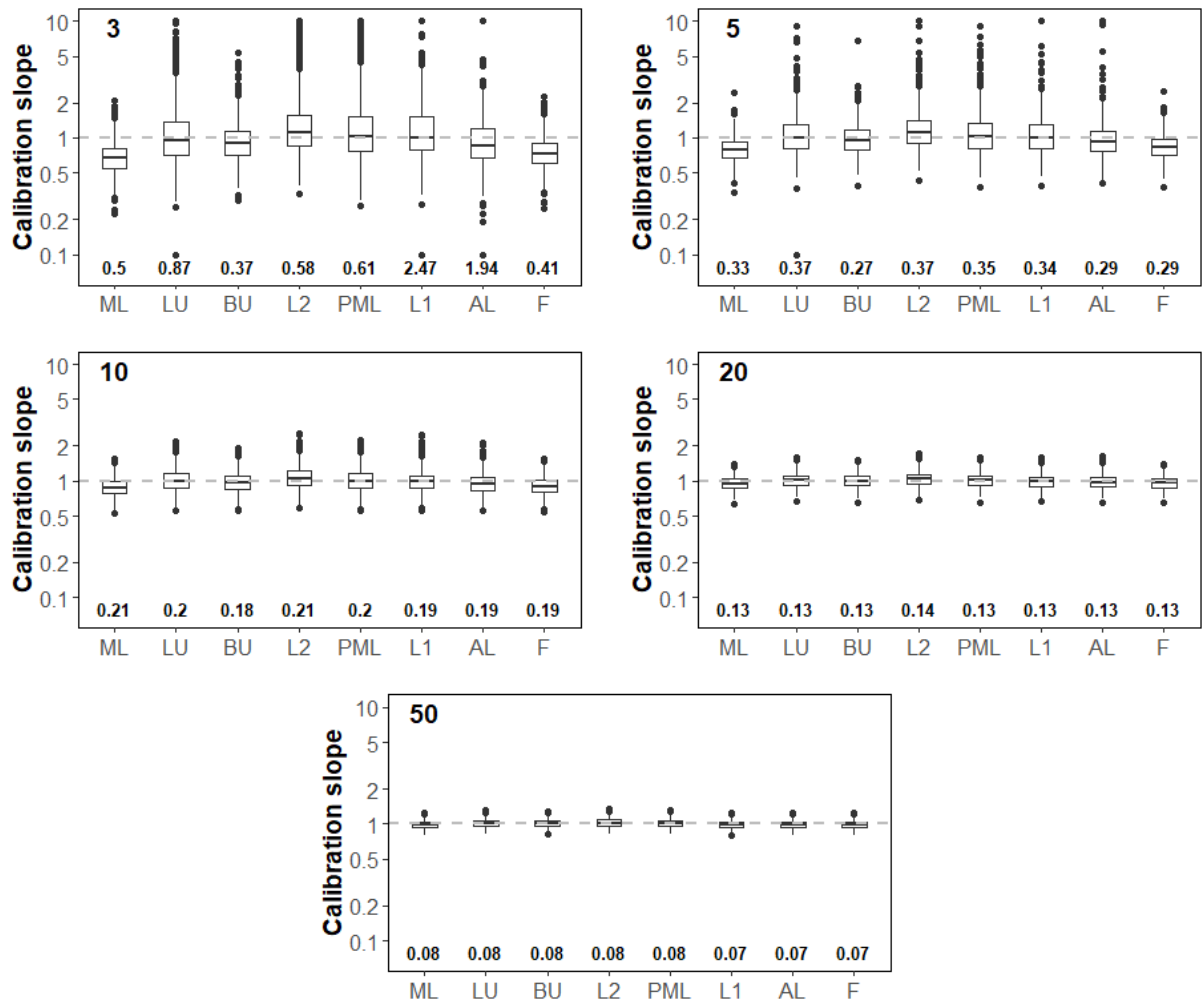
C. Scenarios with 10 true predictors



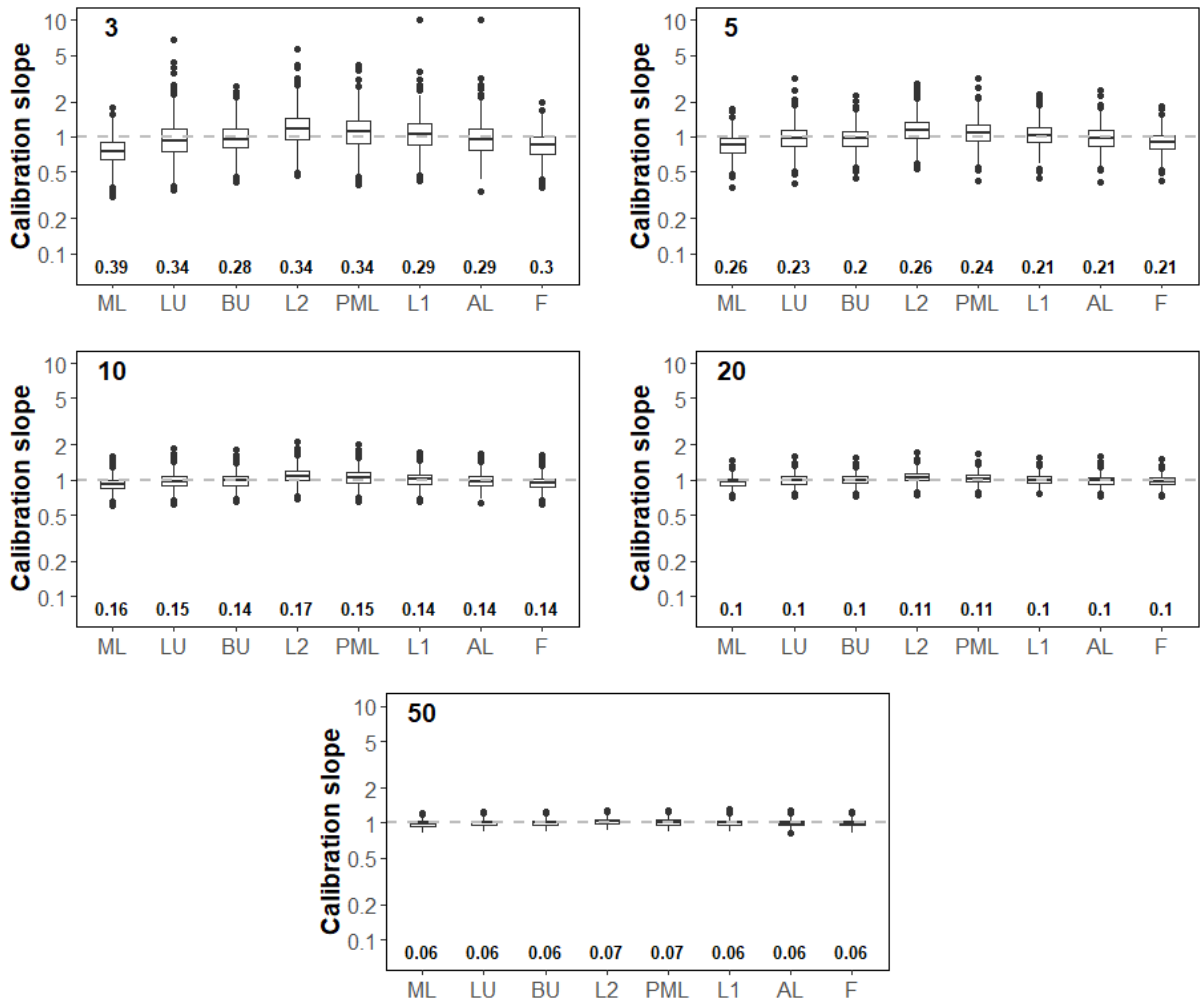
ew

Figure S3. Box plots of calibration slopes over the 1,000 simulation runs for each scenario. The events per variable (EPV) is indicated in the top left. The numbers at the bottom are the root mean squared distances (RMSD) of the log of the calibration slopes. The length of the whiskers is at most 1.5 times the interquartile range. Calibration slopes are winsorized at 0.1 and 10 for visualization purposes. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression ; AL, adaptive LASSO; F, Firth's correction.

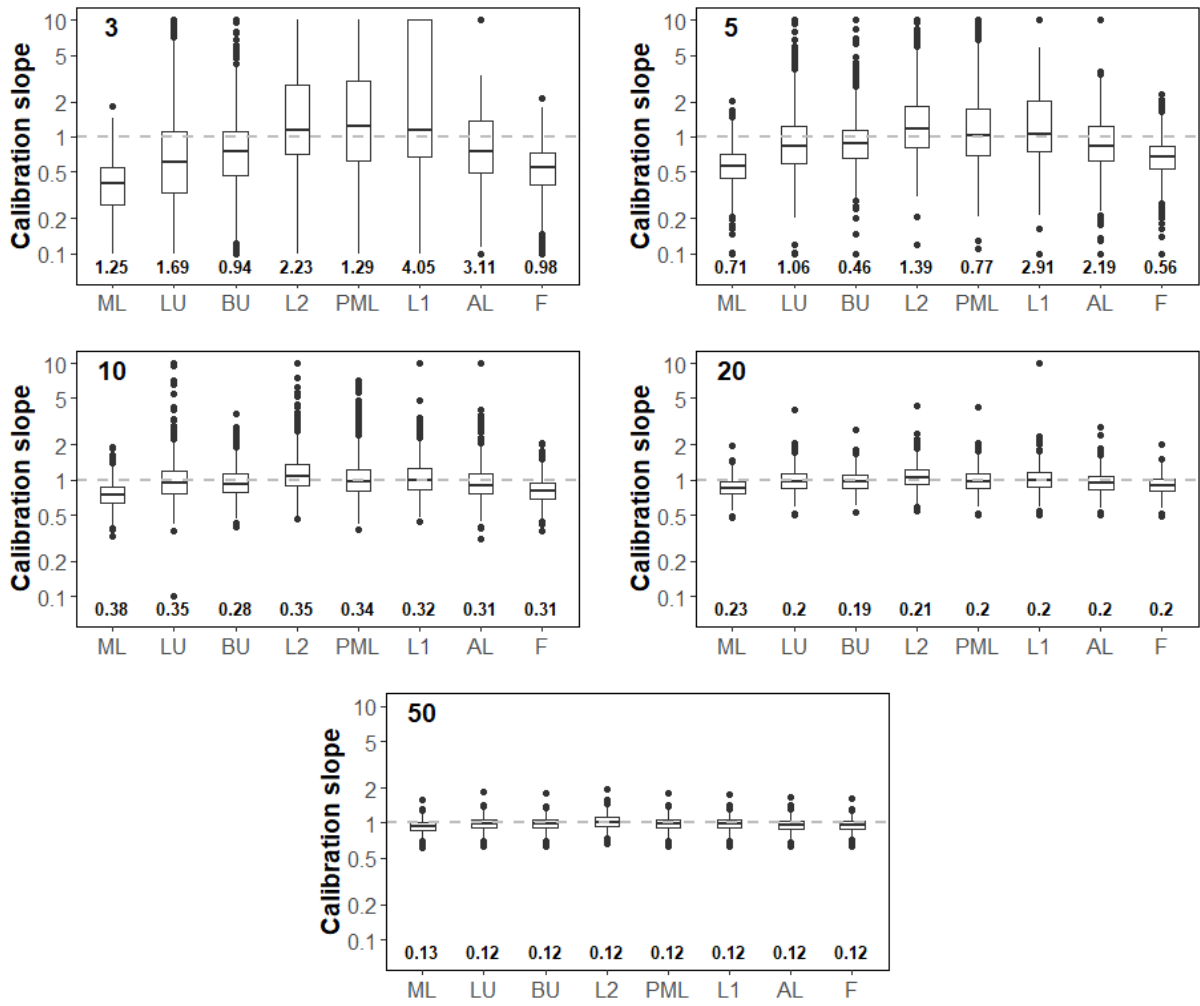
A. 5 true predictors, correlation 0, event rate 10%



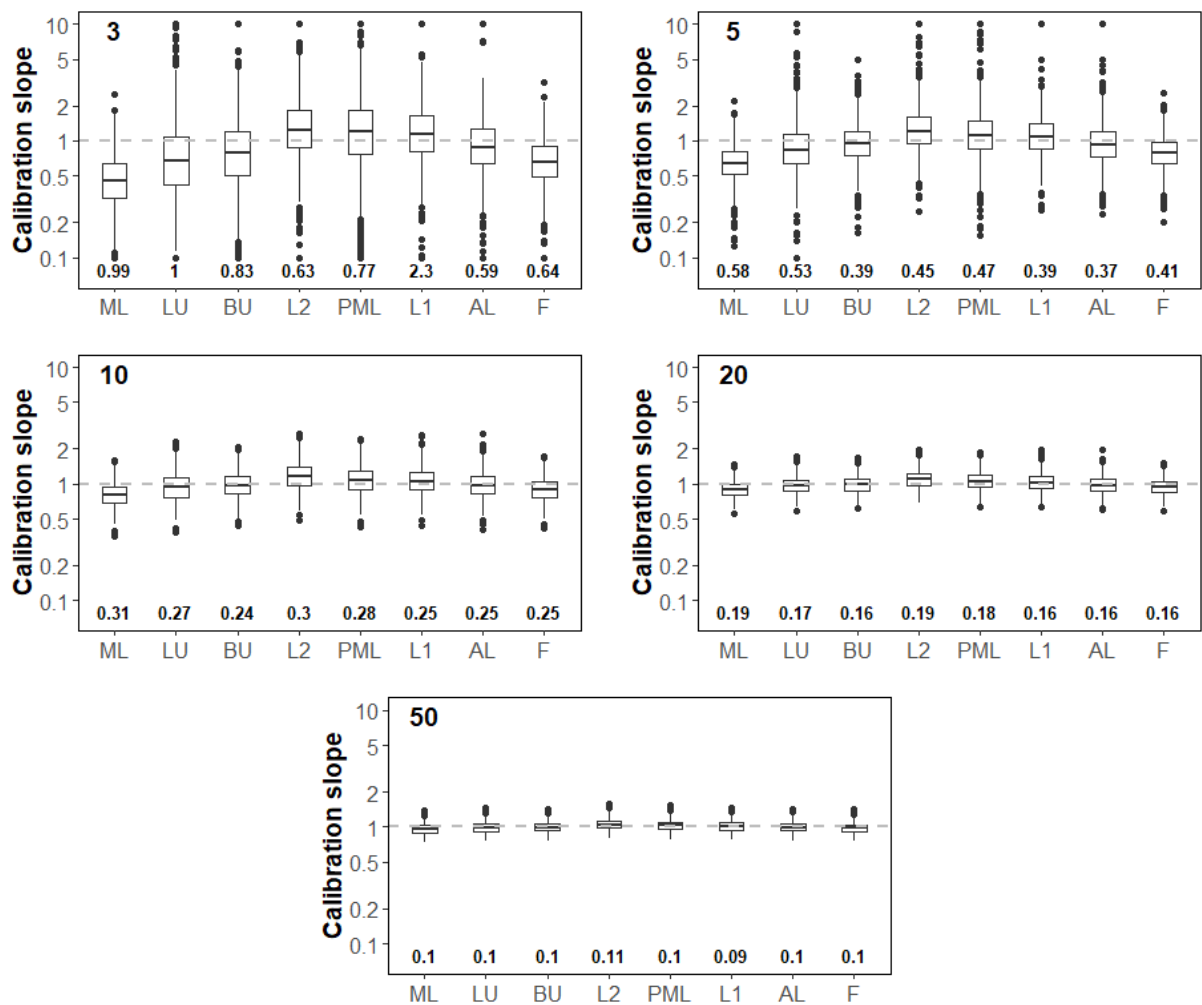
B. 5 true predictors, correlation 0.5, event rate 10%



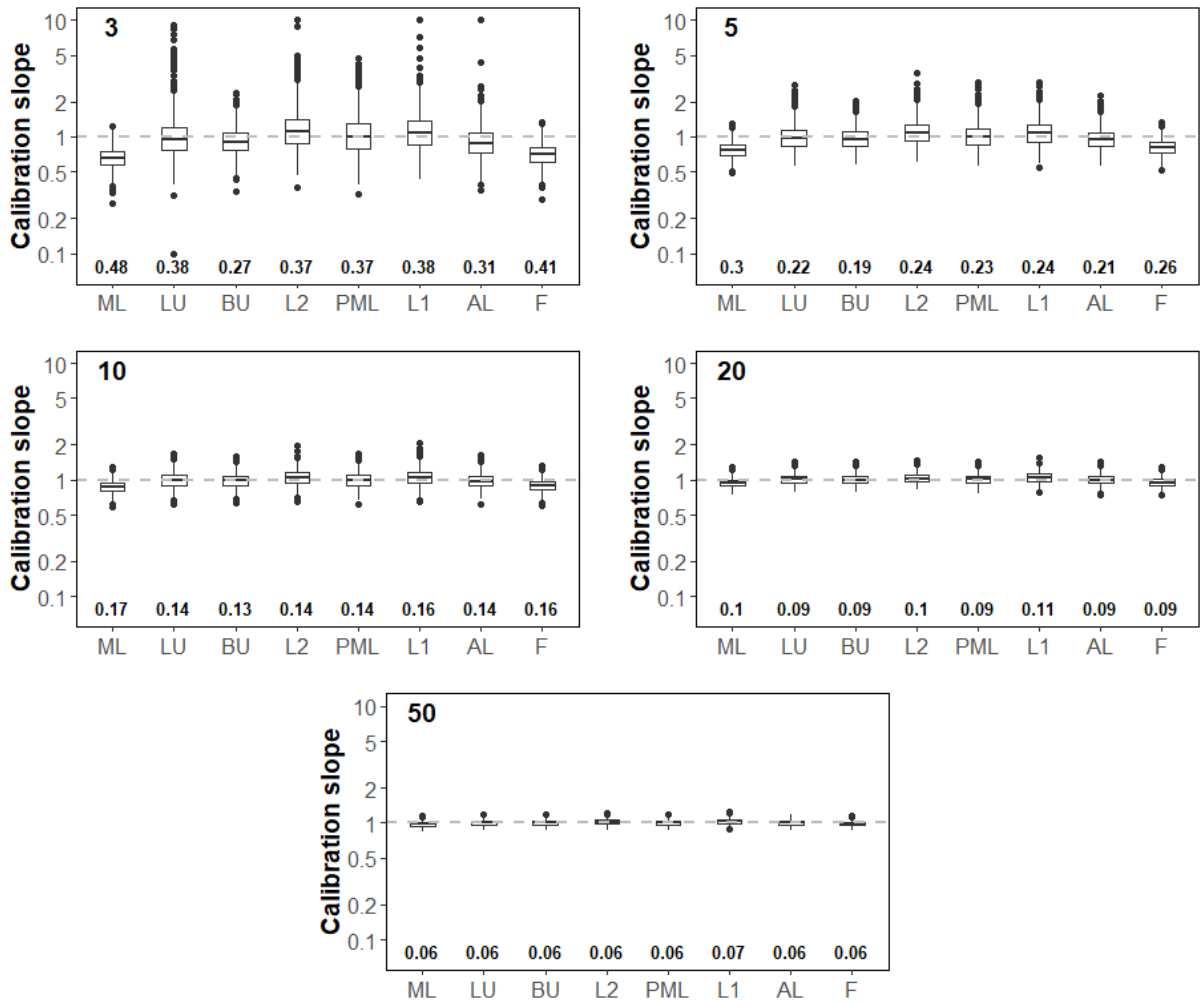
C. 5 true predictors, correlation 0, event rate 50%



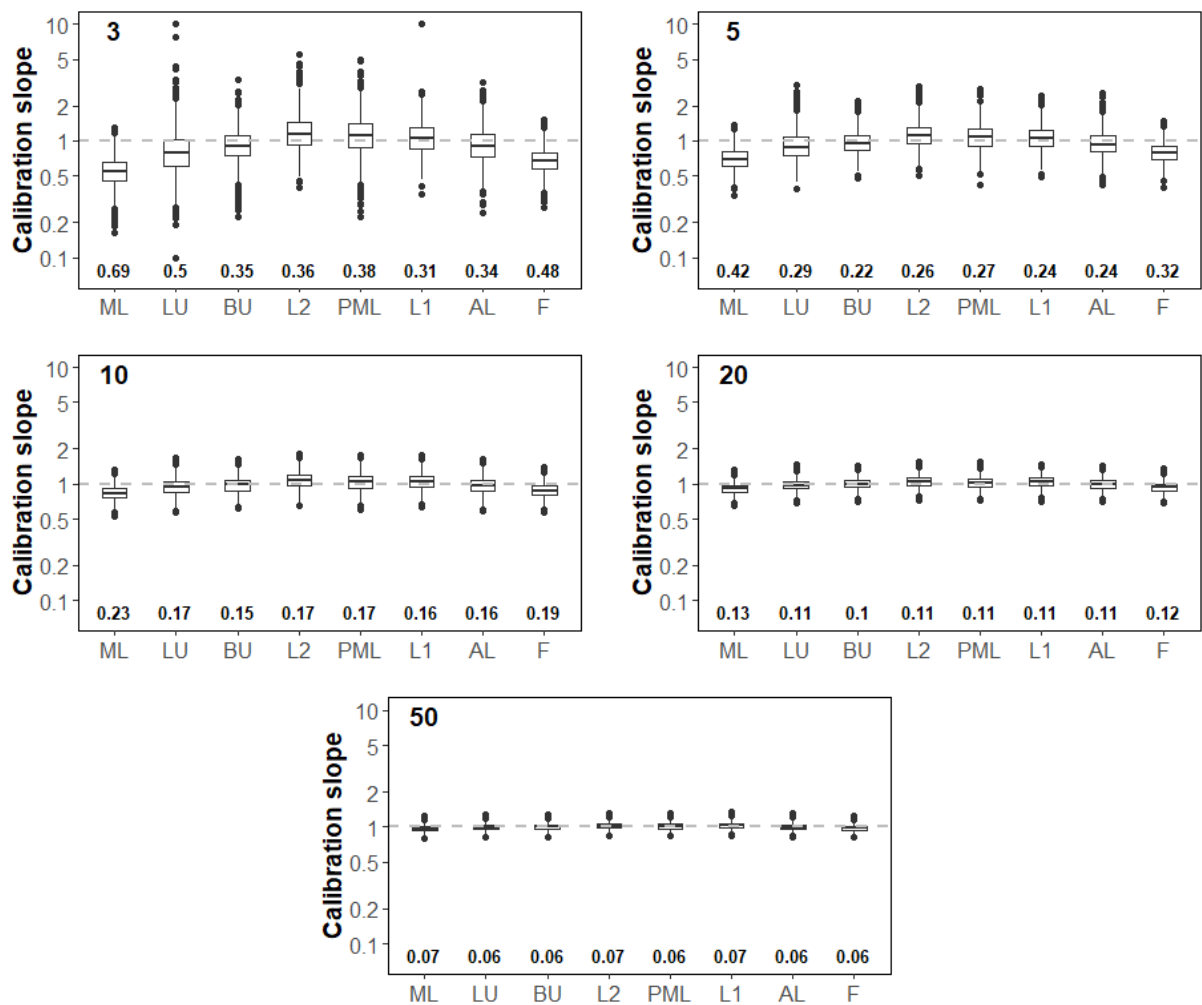
D. 5 true predictors, correlation 0.5, event rate 50%



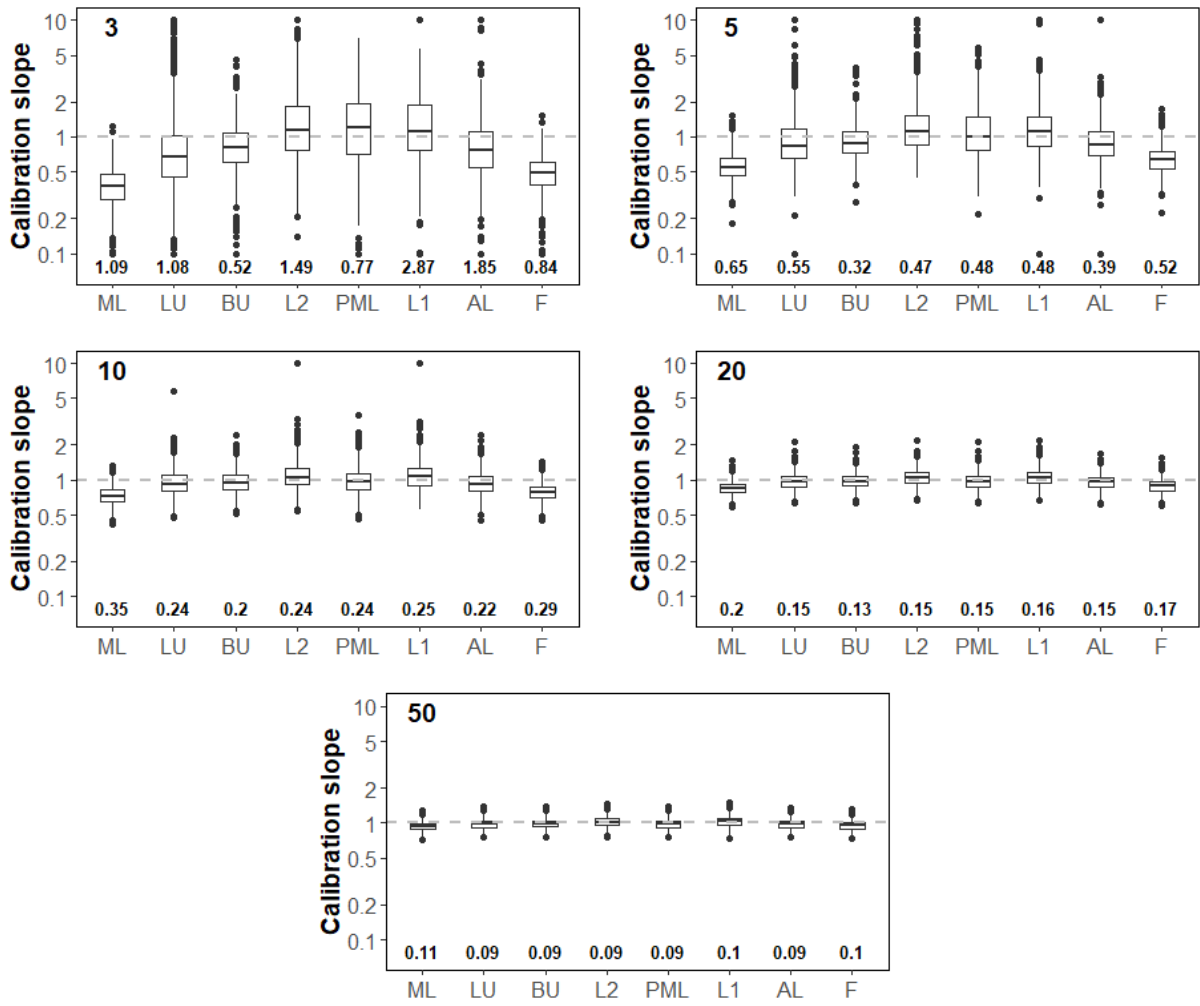
E. 5 true and 5 noise predictors, correlation 0, event rate 10%



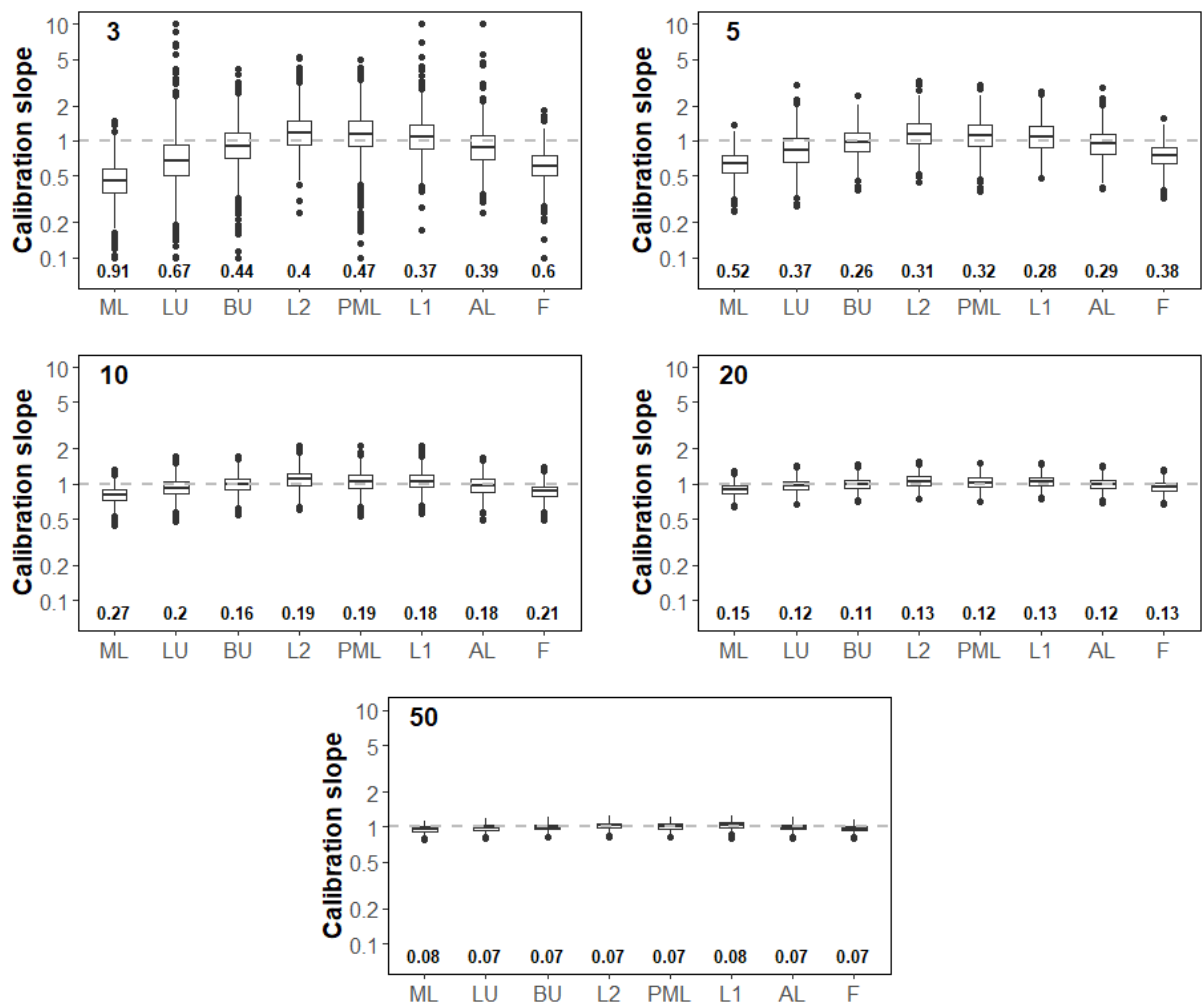
F. 5 true and 5 noise predictors, correlation 0.5, event rate 10%



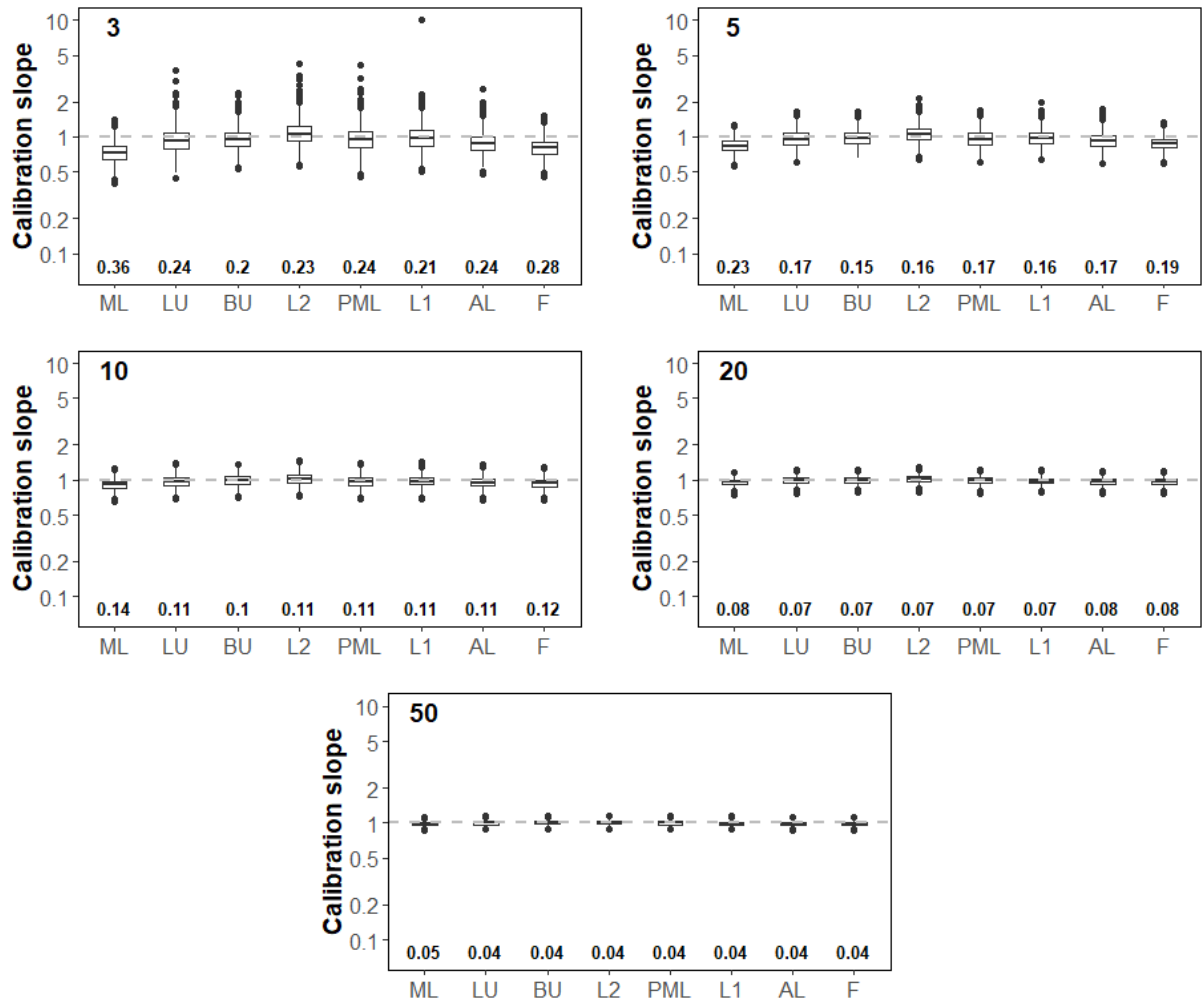
G. 5 true and 5 noise predictors, correlation 0, event rate 50%



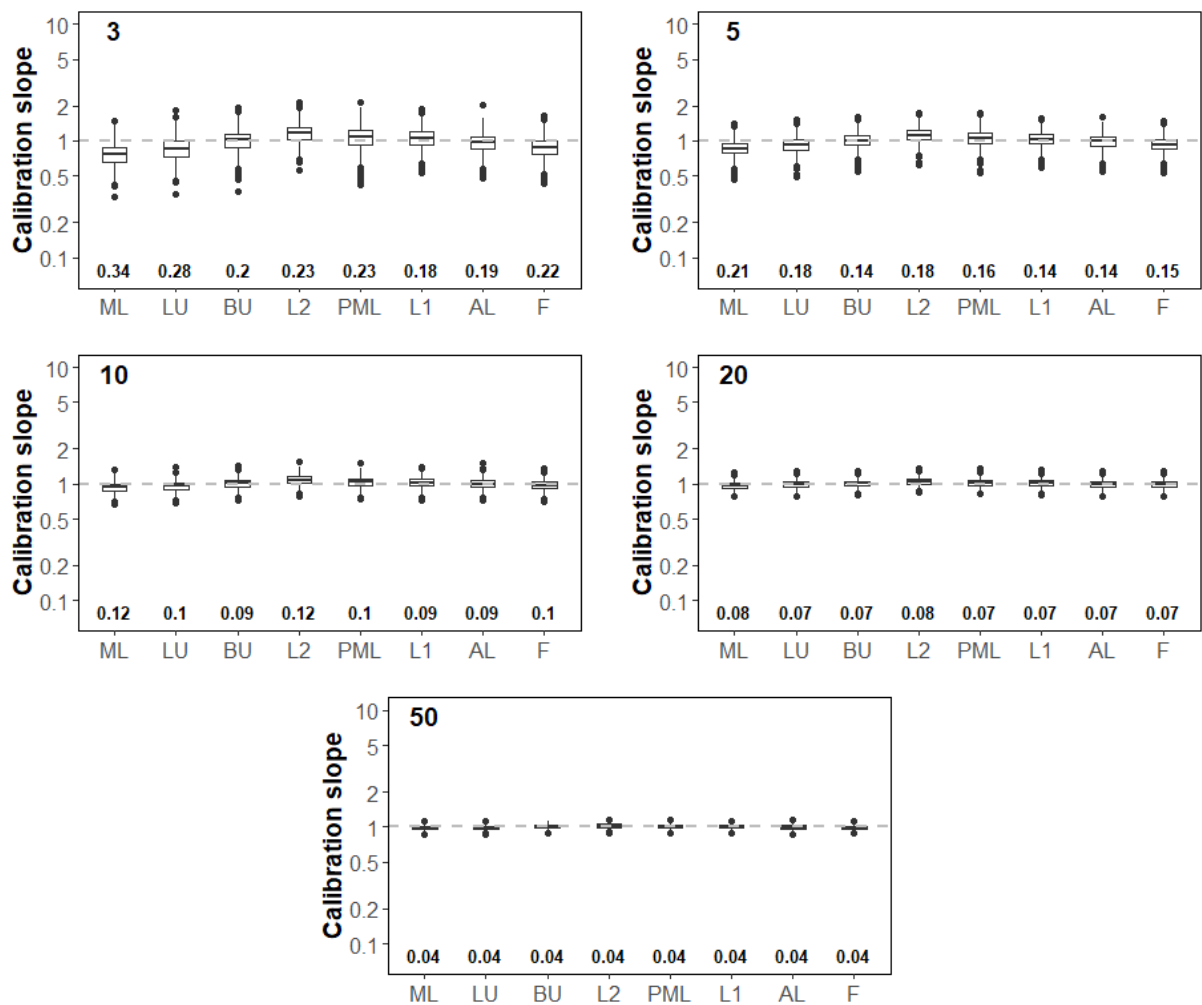
H. 5 true and 5 noise predictors, correlation 0.5, event rate 50%



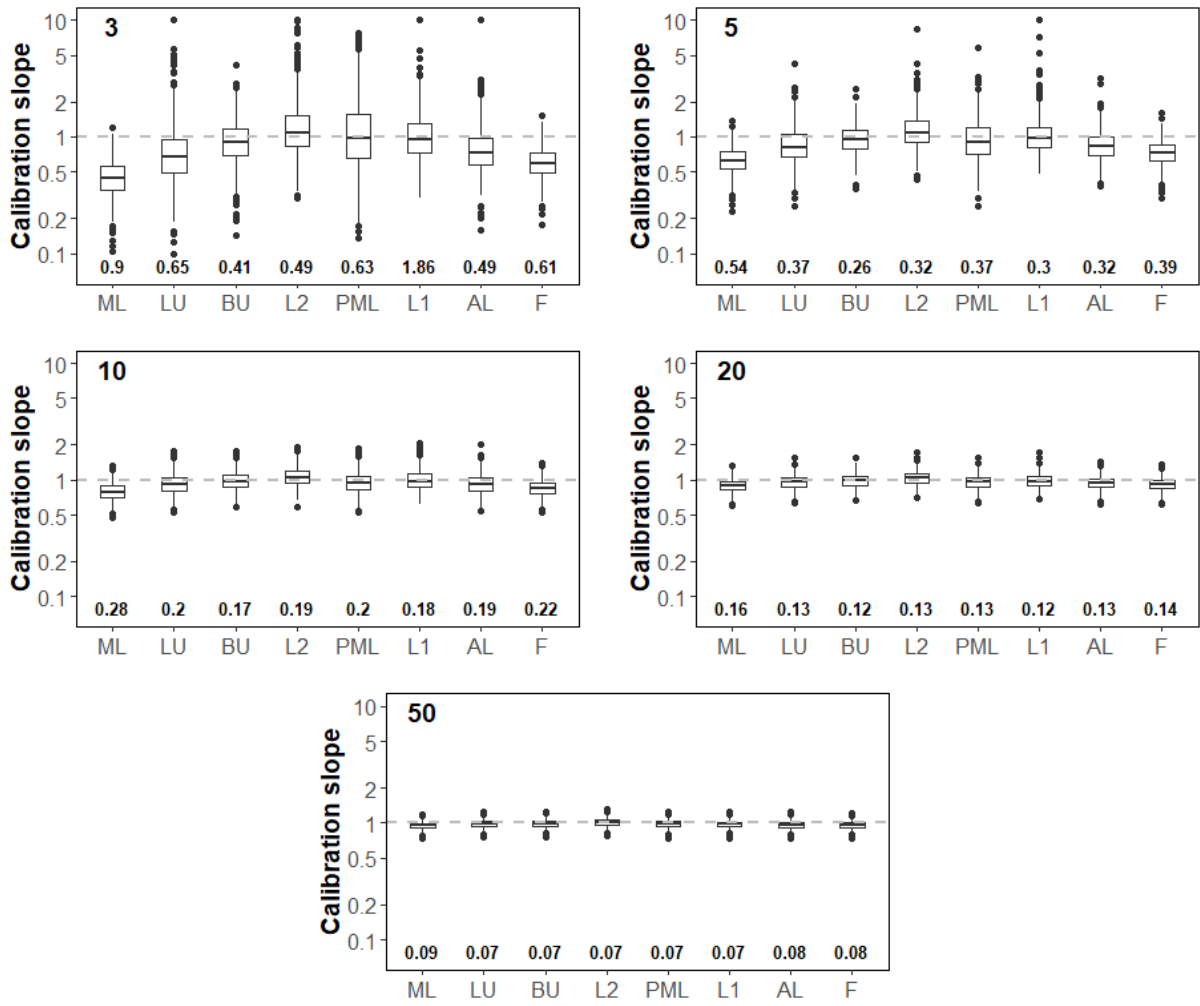
I. 10 true predictors, correlation 0, event rate 10%



J. 10 true predictors, correlation 0.5, event rate 10%



K. 10 true predictors, correlation 0, event rate 50%



L. 10 true predictors, correlation 0.5, event rate 50%

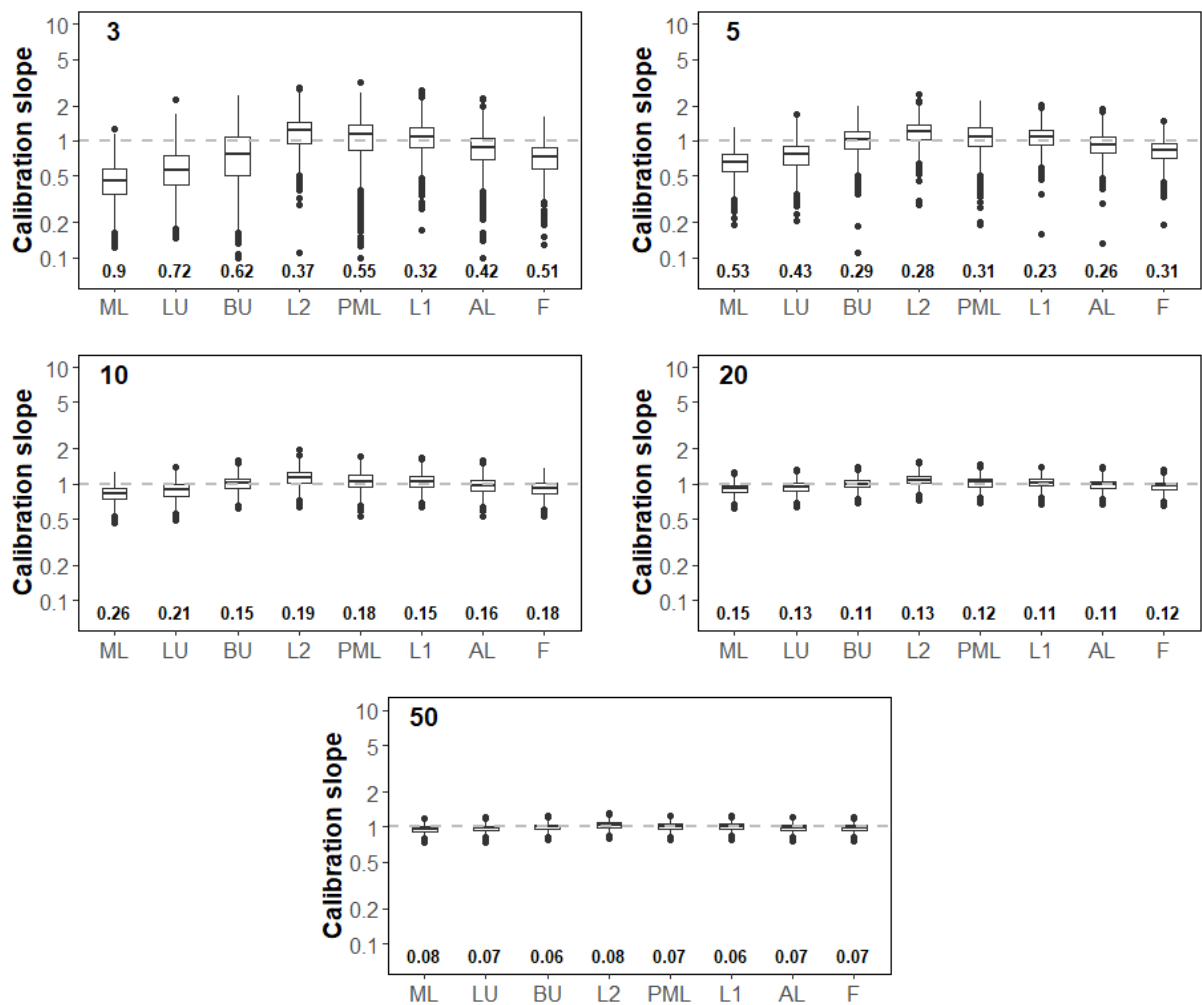
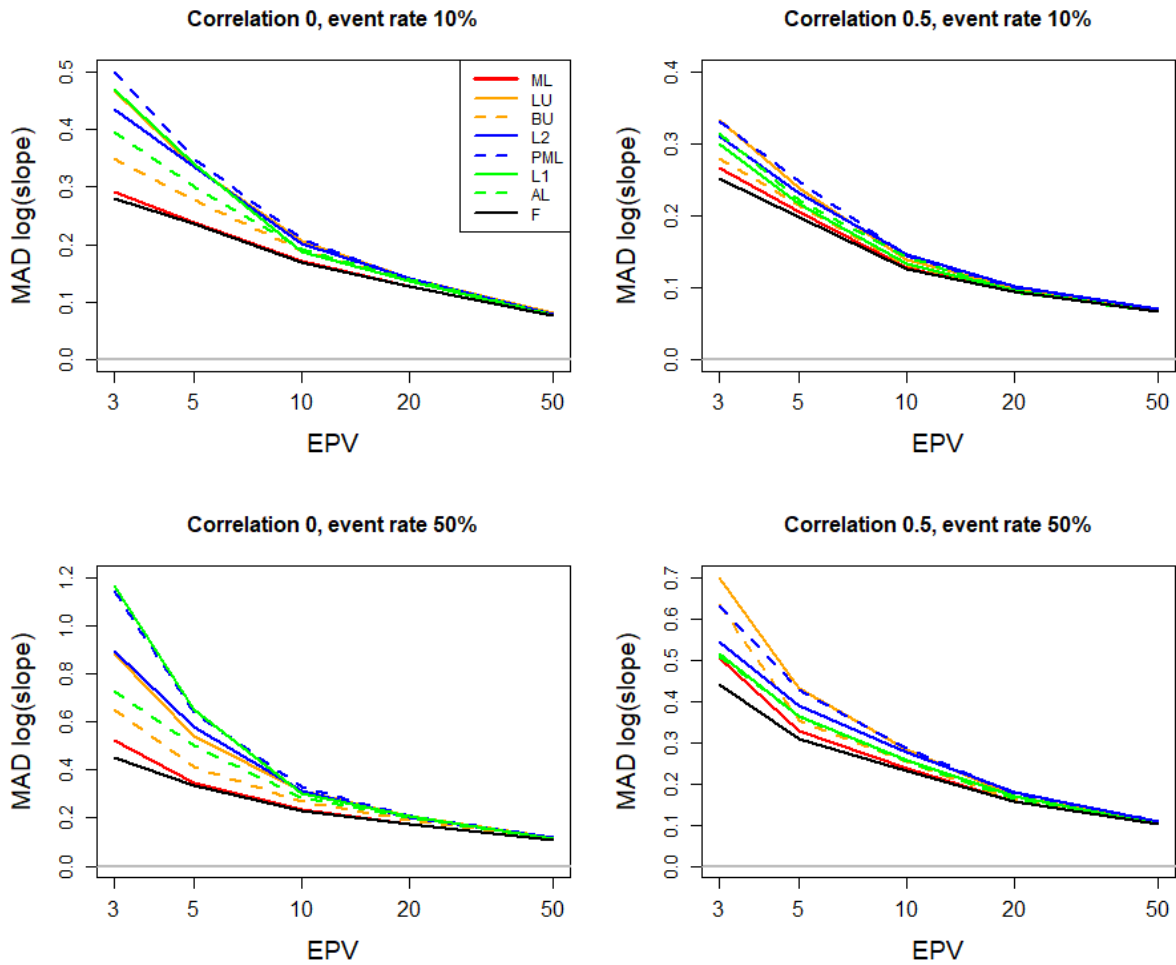
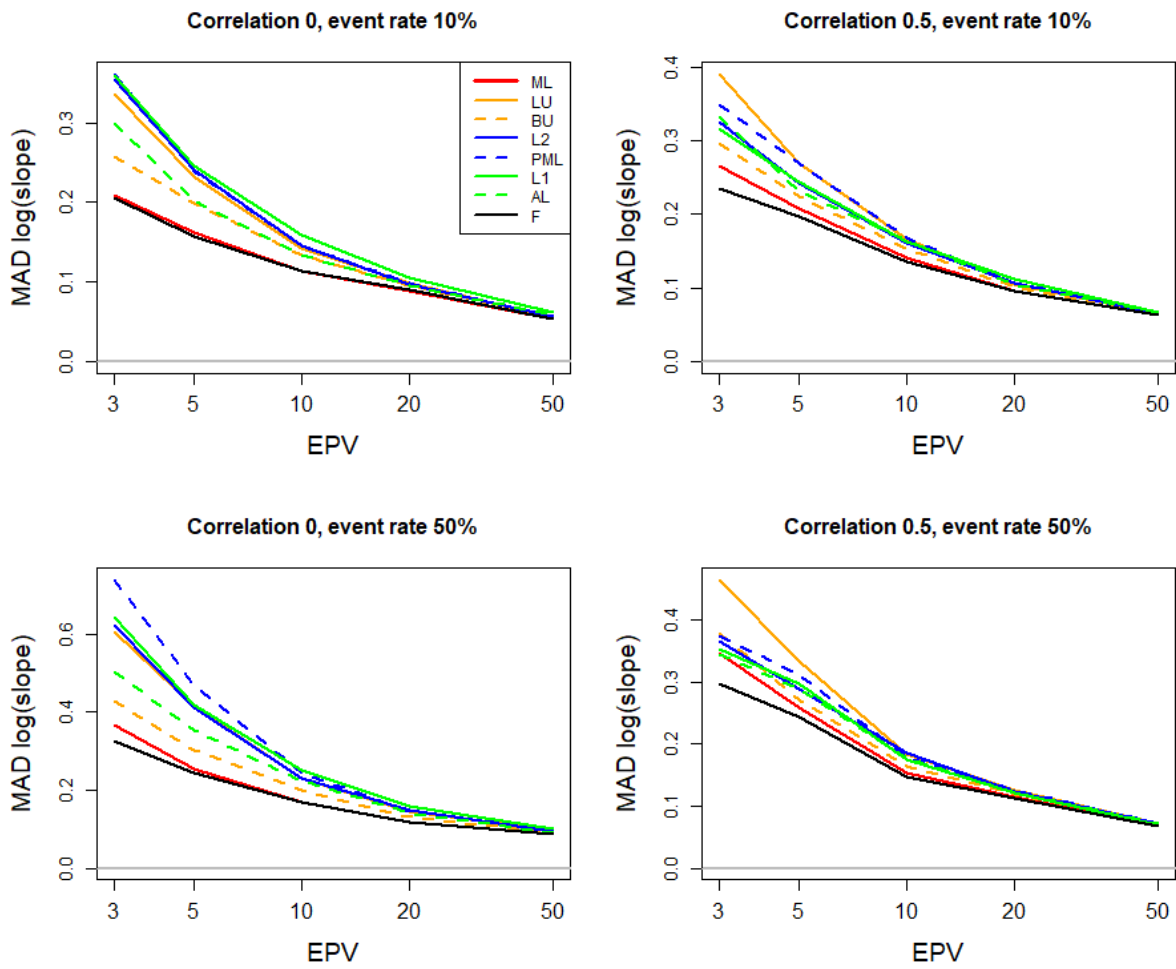


Figure S4. Median absolute deviation (MAD) of the logarithm of the calibration slope. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression ; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

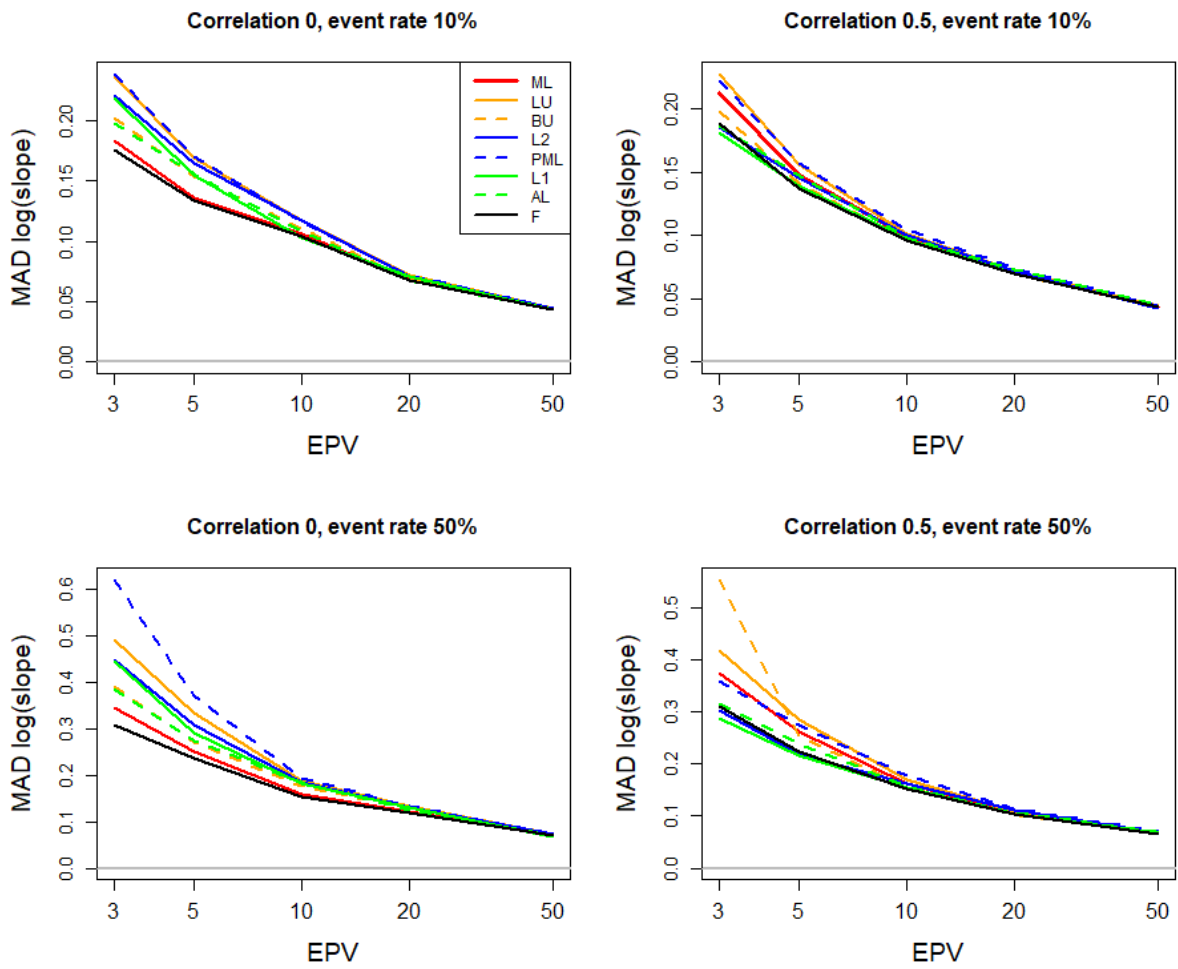


B. Scenarios with 5 true and 5 noise predictors



ew

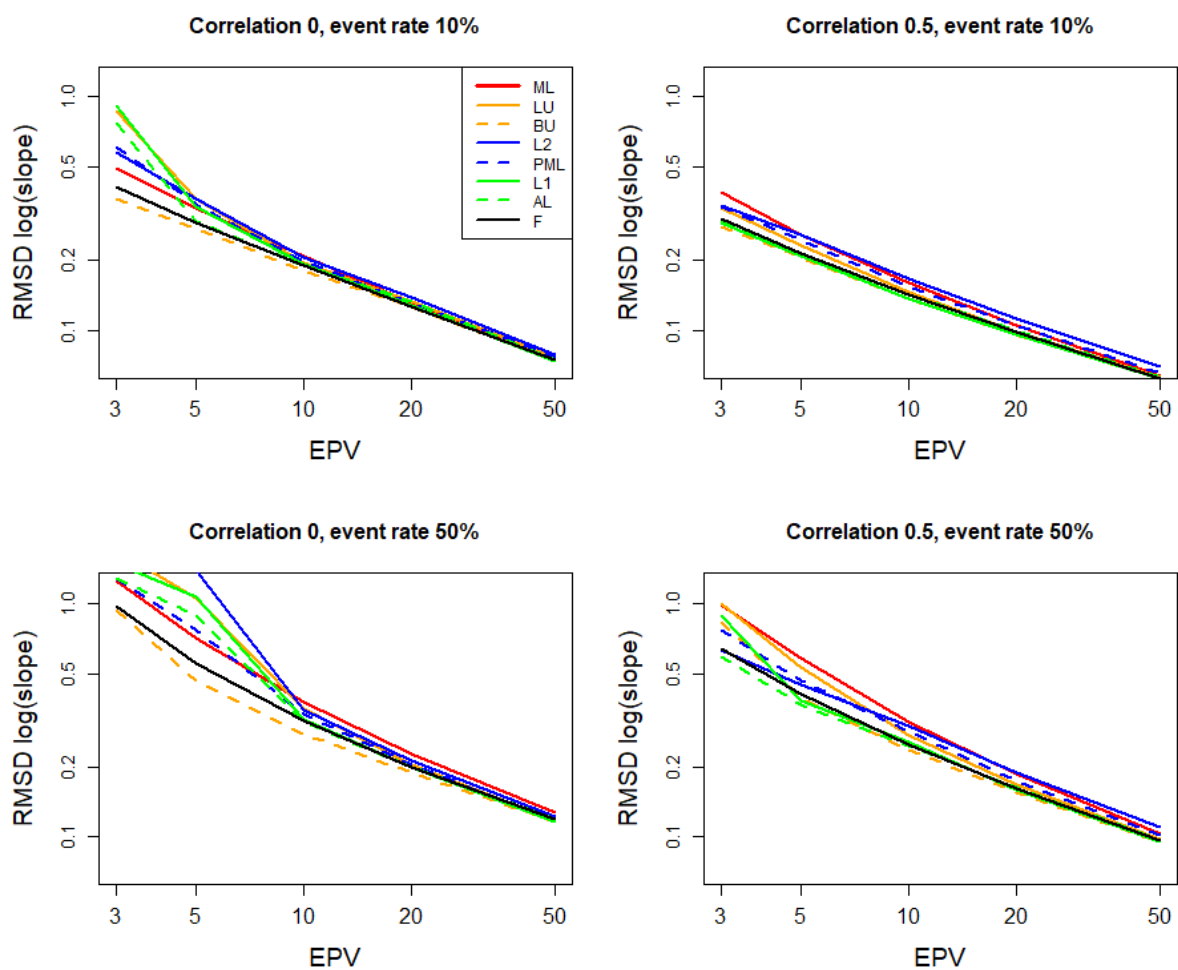
C. Scenarios with 10 true predictors



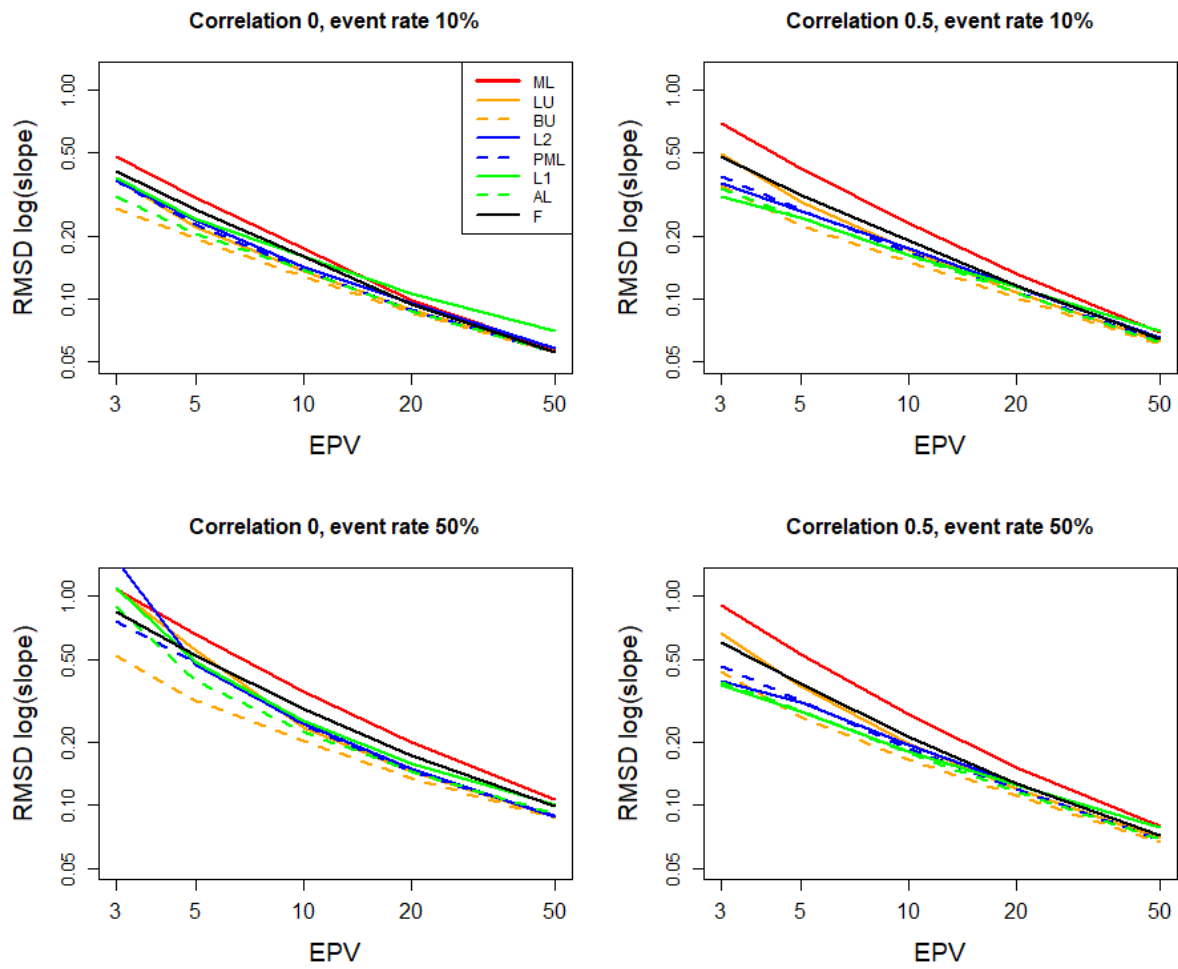
ew

Figure S5. Root mean squared distance of the target value (RMSD) of the calibration slope. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

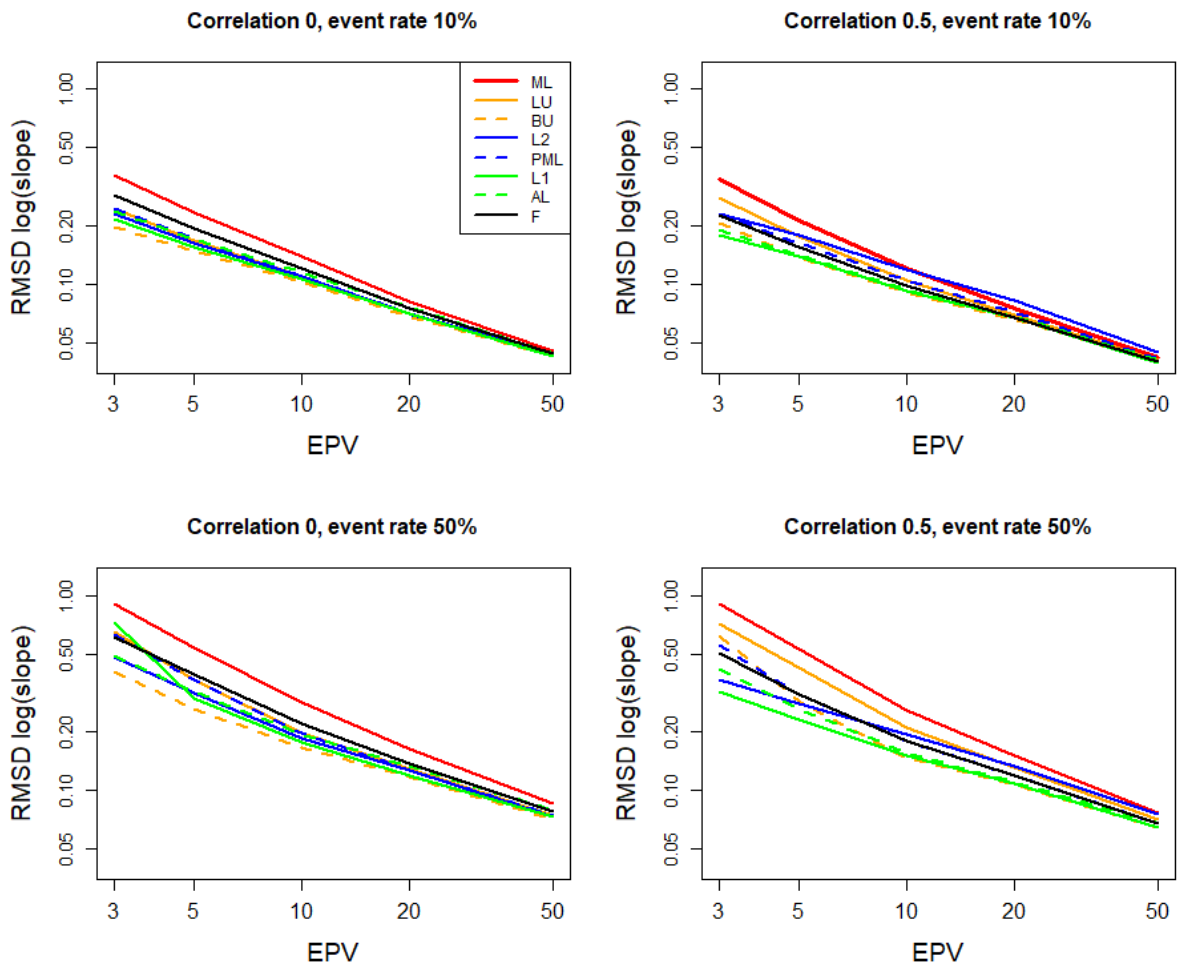


B. Scenarios with 5 true and 5 noise predictors



ew

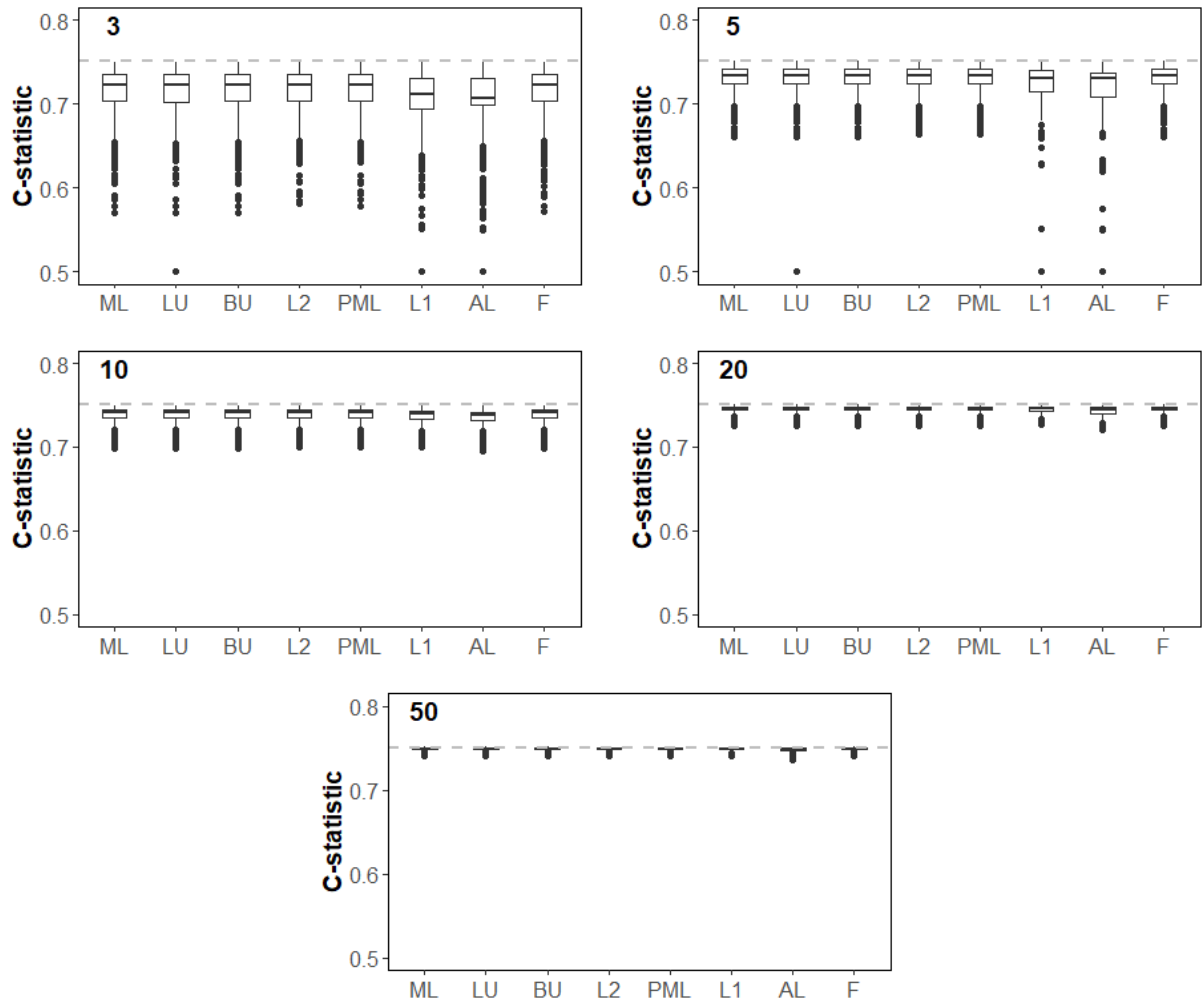
C. Scenarios with 10 true predictors



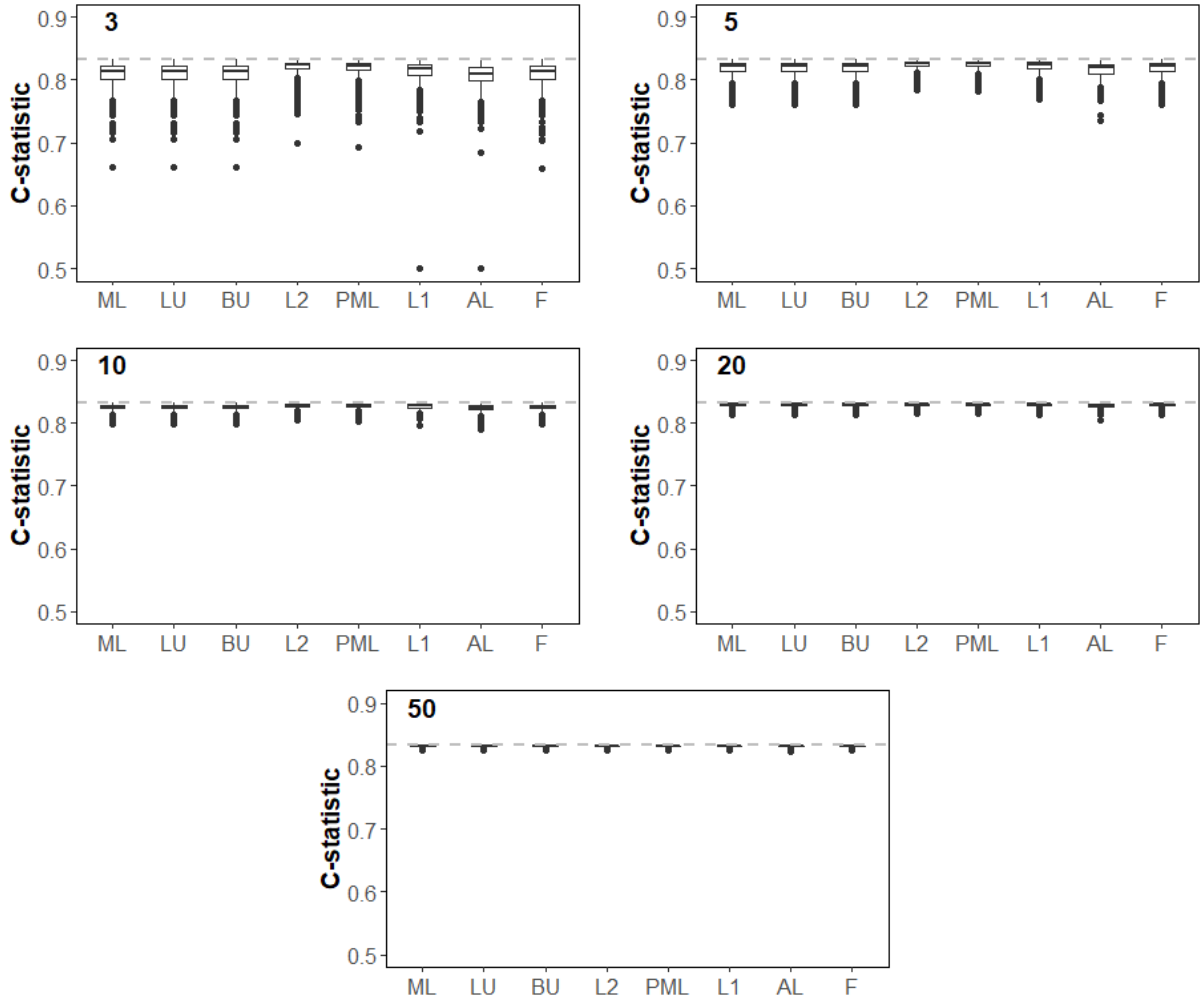
ew

Figure S6. Box plots of c-statistics over the 1,000 simulation runs for each scenario. The events per variable (EPV) is indicated in the top left. The length of the whiskers is at most 1.5 times the interquartile range. C-statistics are winsorized at 0.5 for visualization purposes. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

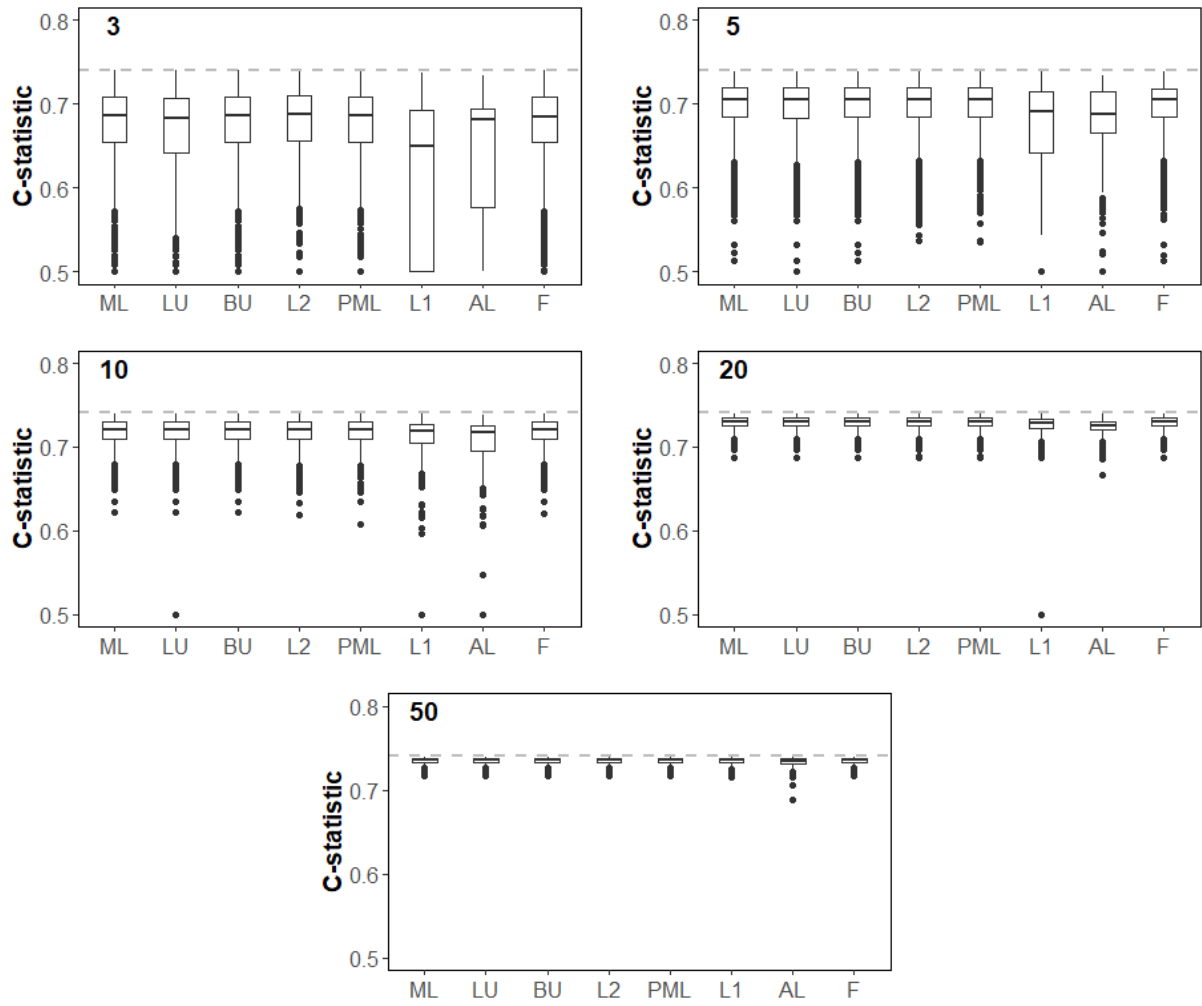
A. 5 true predictors, correlation 0, event rate 10%



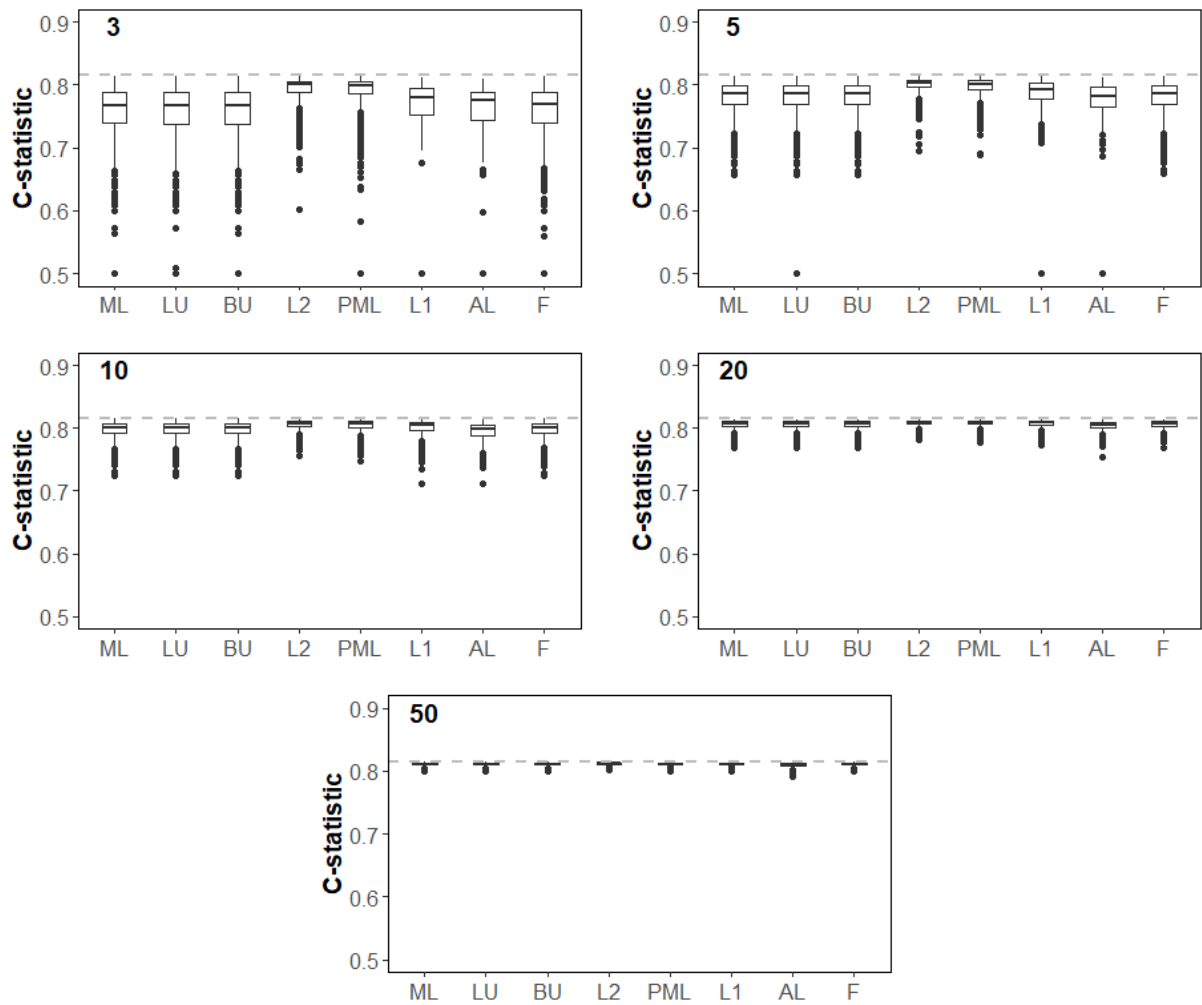
B. 5 true predictors, correlation 0.5, event rate 10%



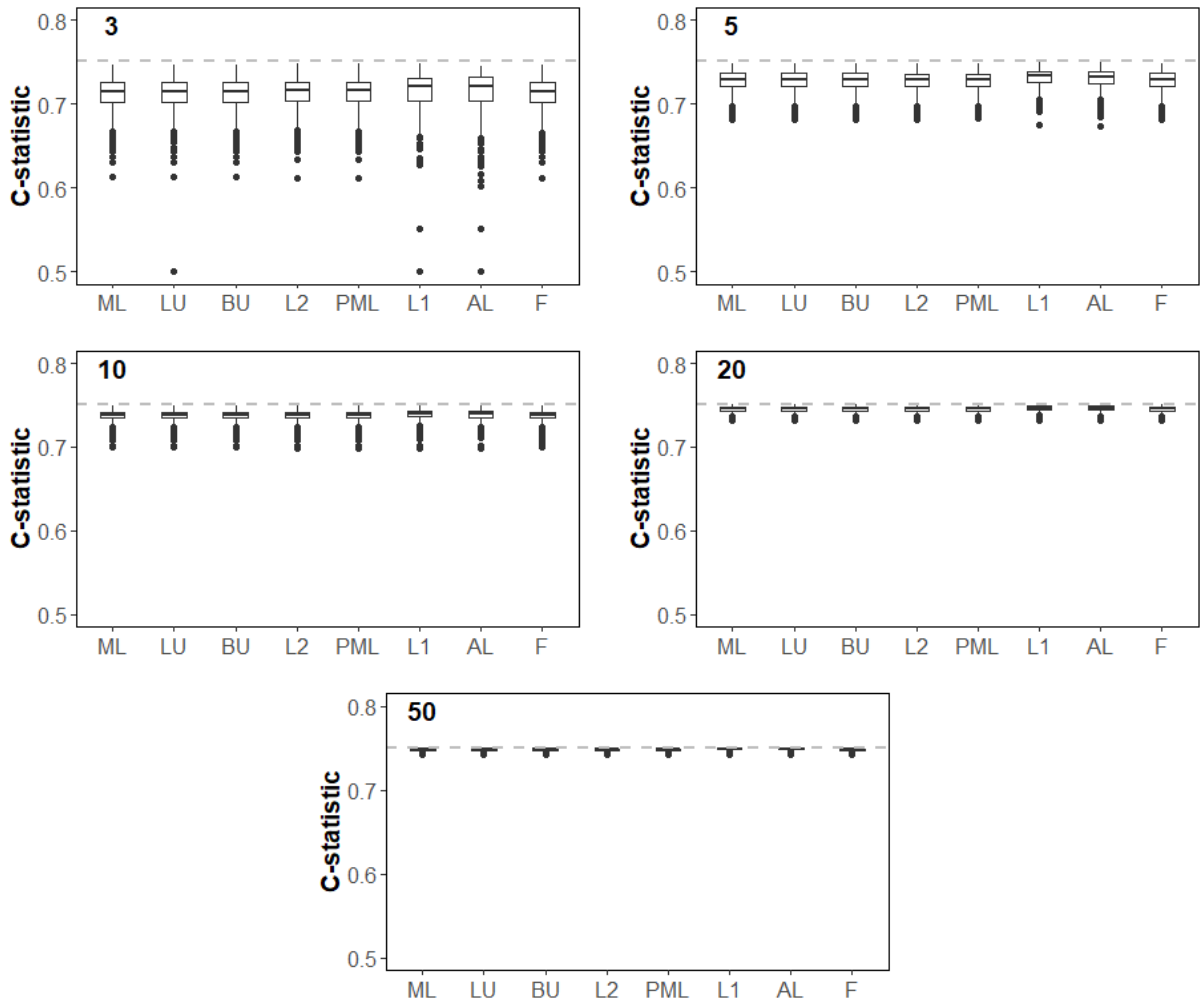
C. 5 true predictors, correlation 0, event rate 50%



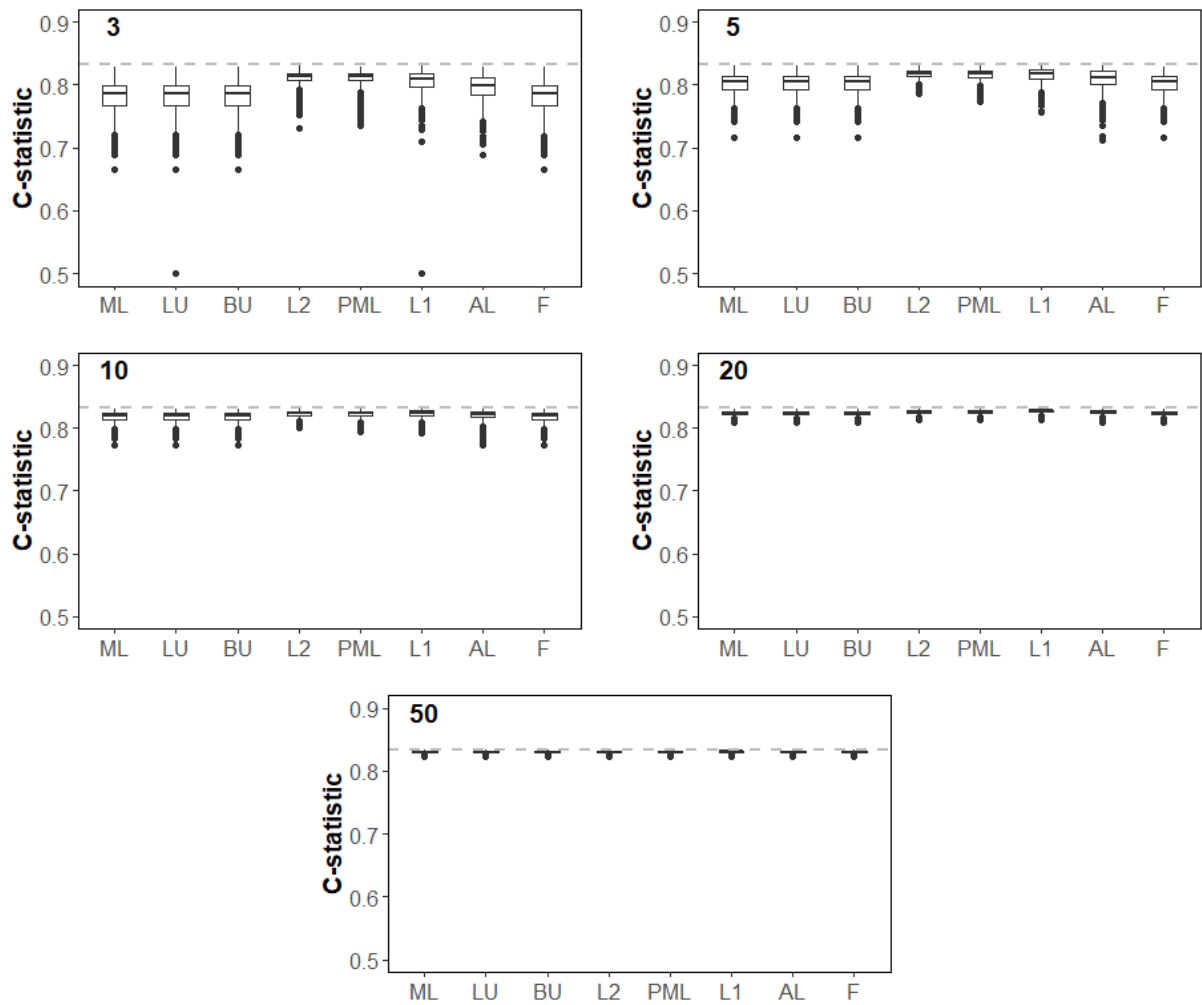
D. 5 true predictors, correlation 0.5, event rate 50%



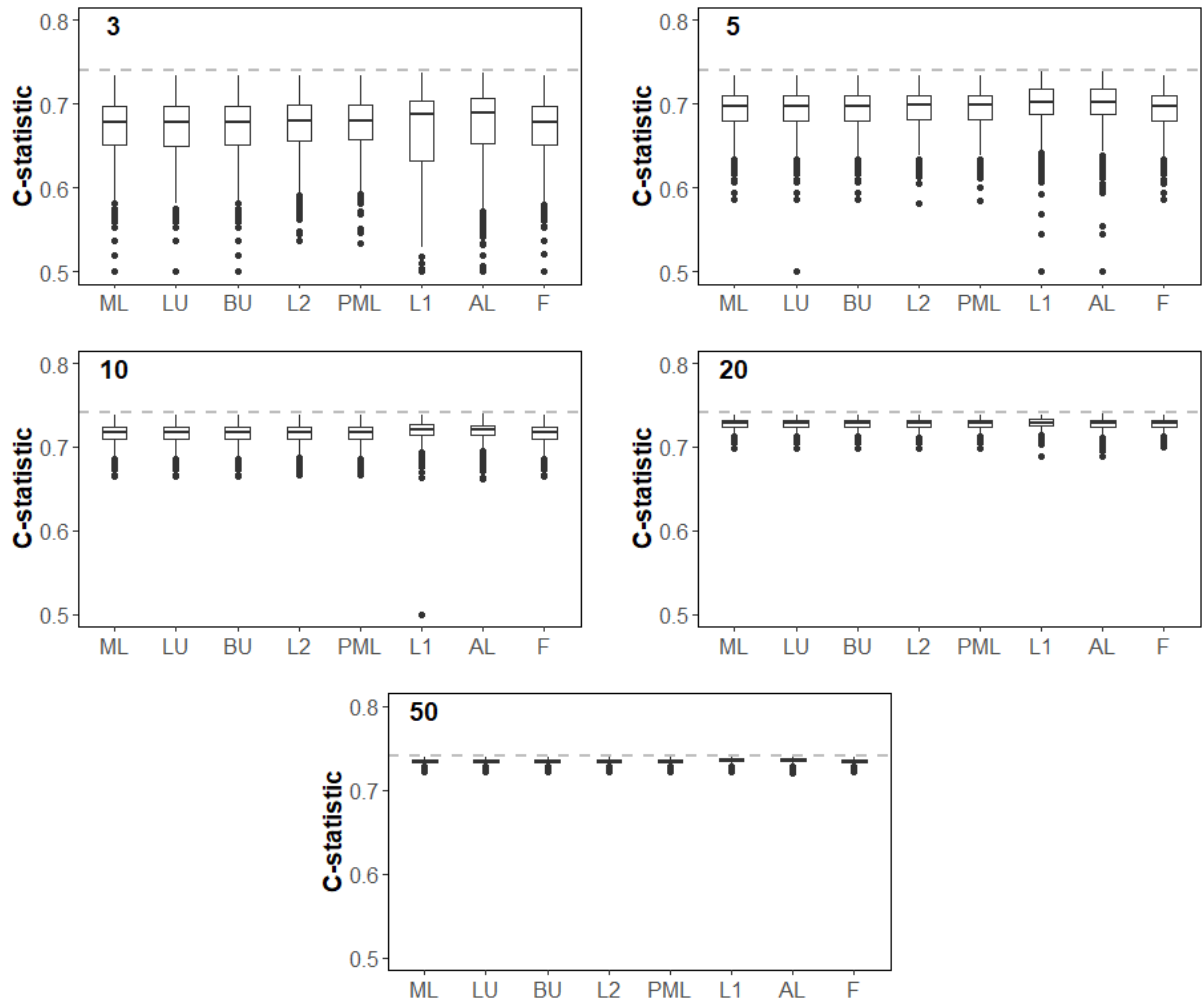
E. 5 true and 5 noise predictors, correlation 0, event rate 10%



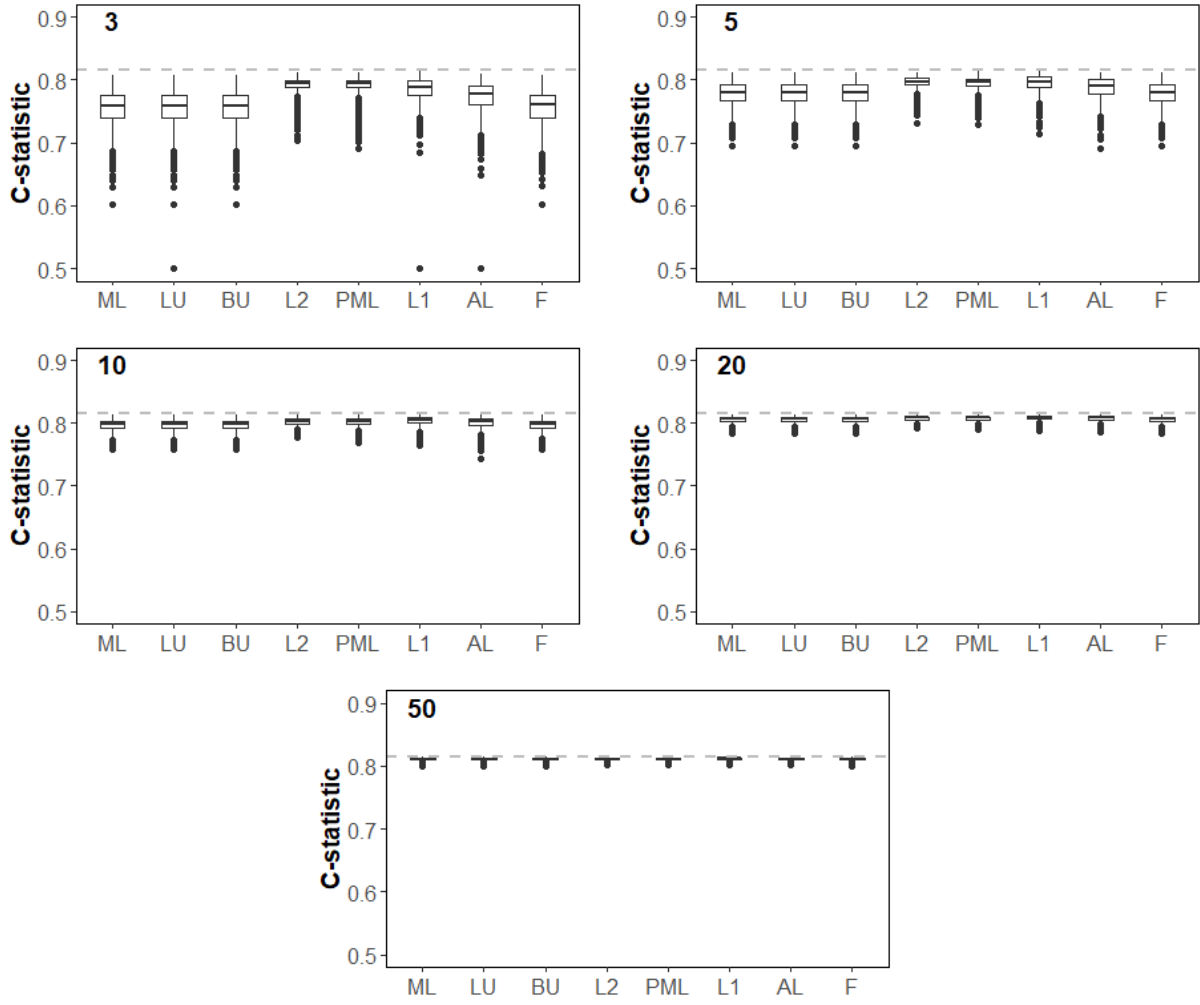
F. 5 true and 5 noise predictors, correlation 0.5, event rate 10%



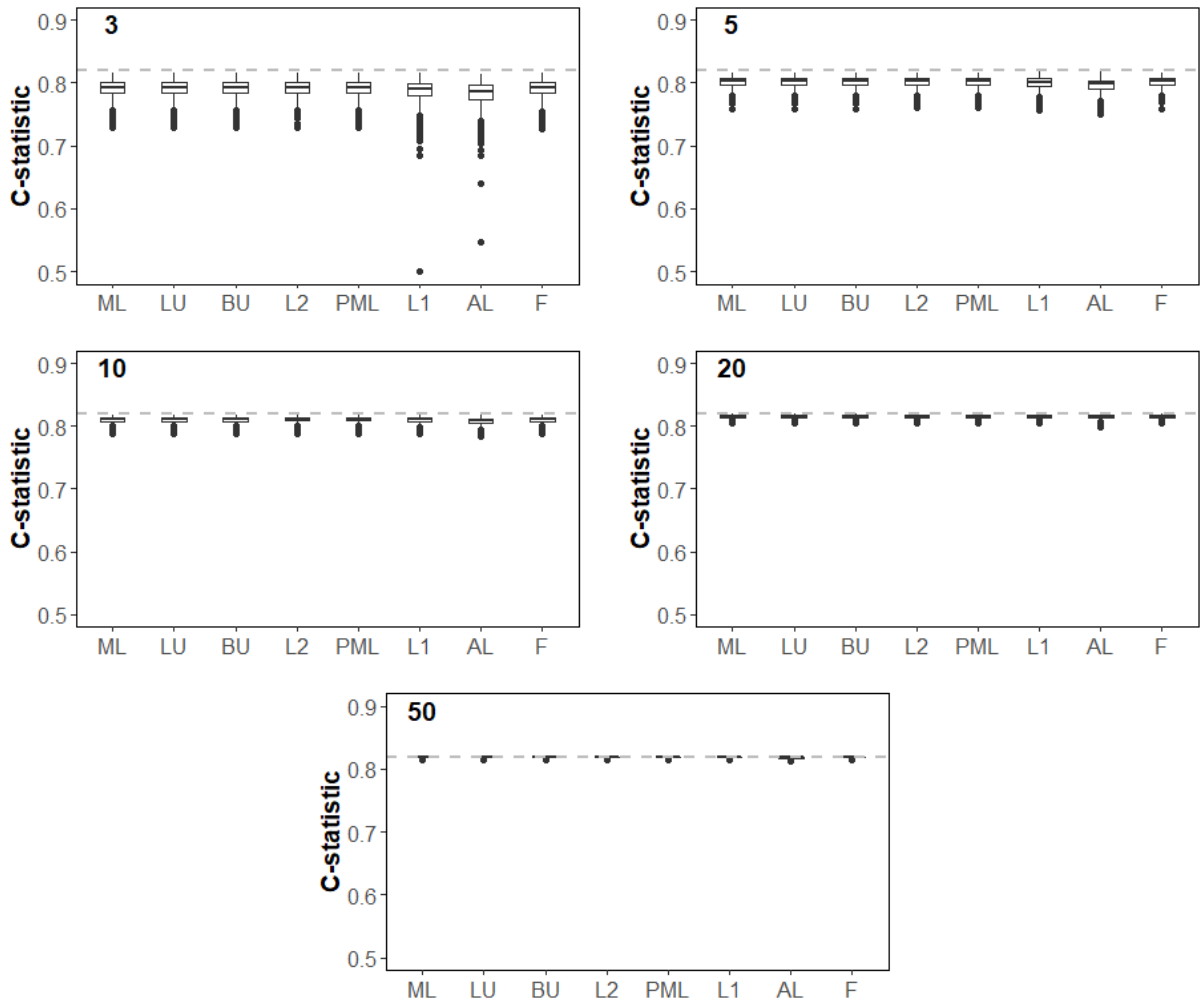
G. 5 true and 5 noise predictors, correlation 0, event rate 50%



H. 5 true and 5 noise predictors, correlation 0.5, event rate 50%

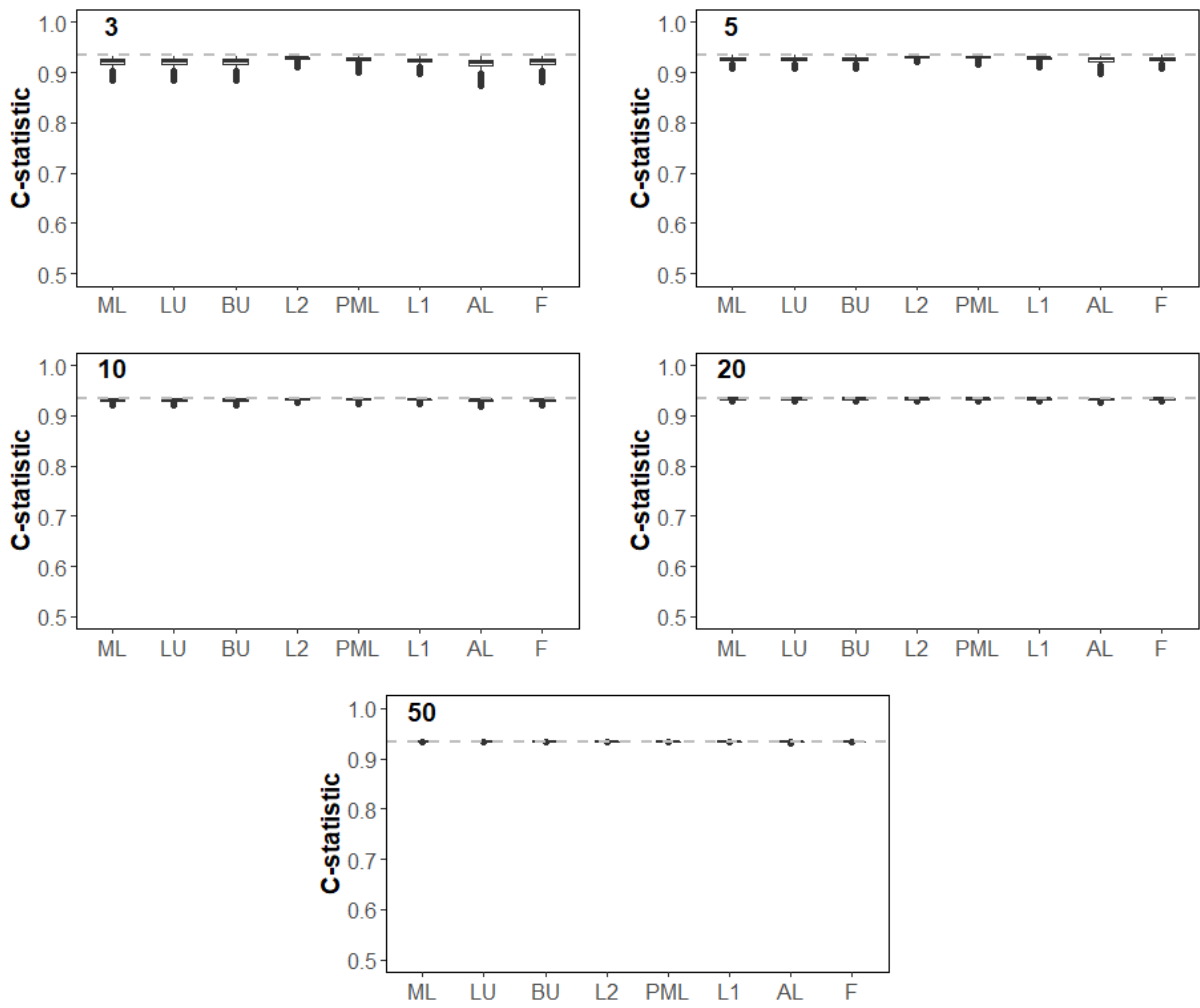


I. 10 true predictors, correlation 0, event rate 10%



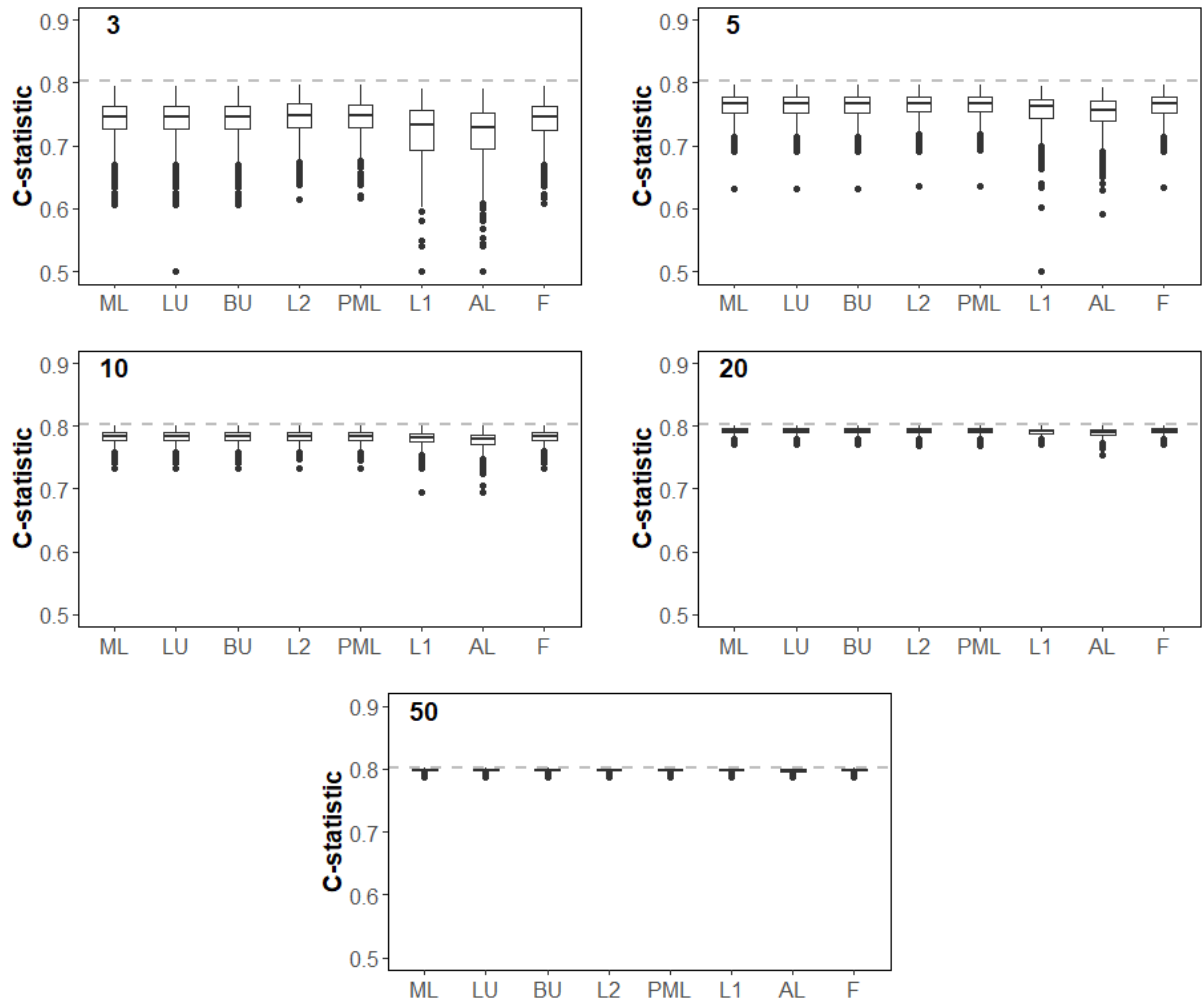
ew

J. 10 true predictors, correlation 0.5, event rate 10%



ew

K. 10 true predictors, correlation 0, event rate 50%



L. 10 true predictors, correlation 0.5, event rate 50%

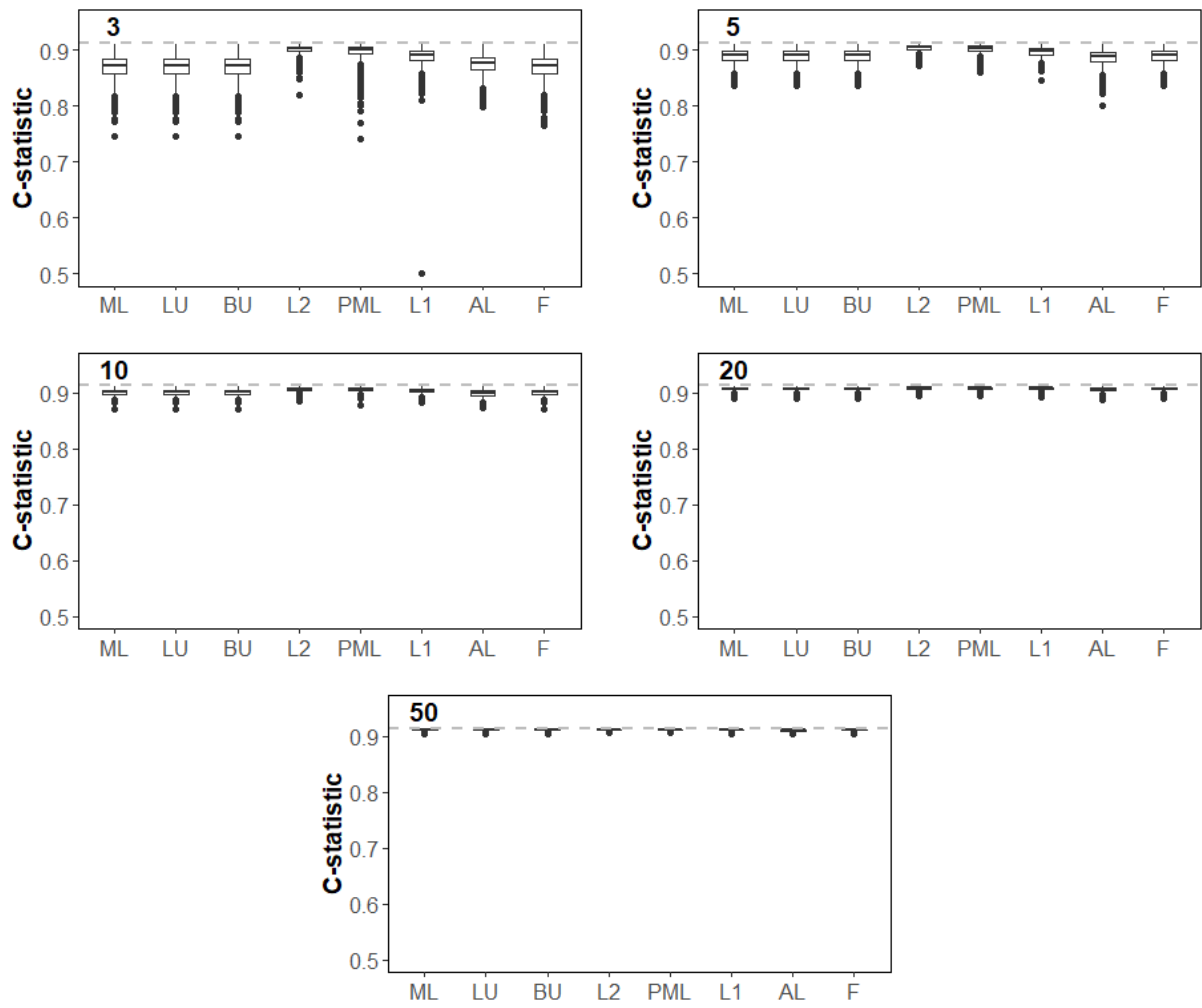
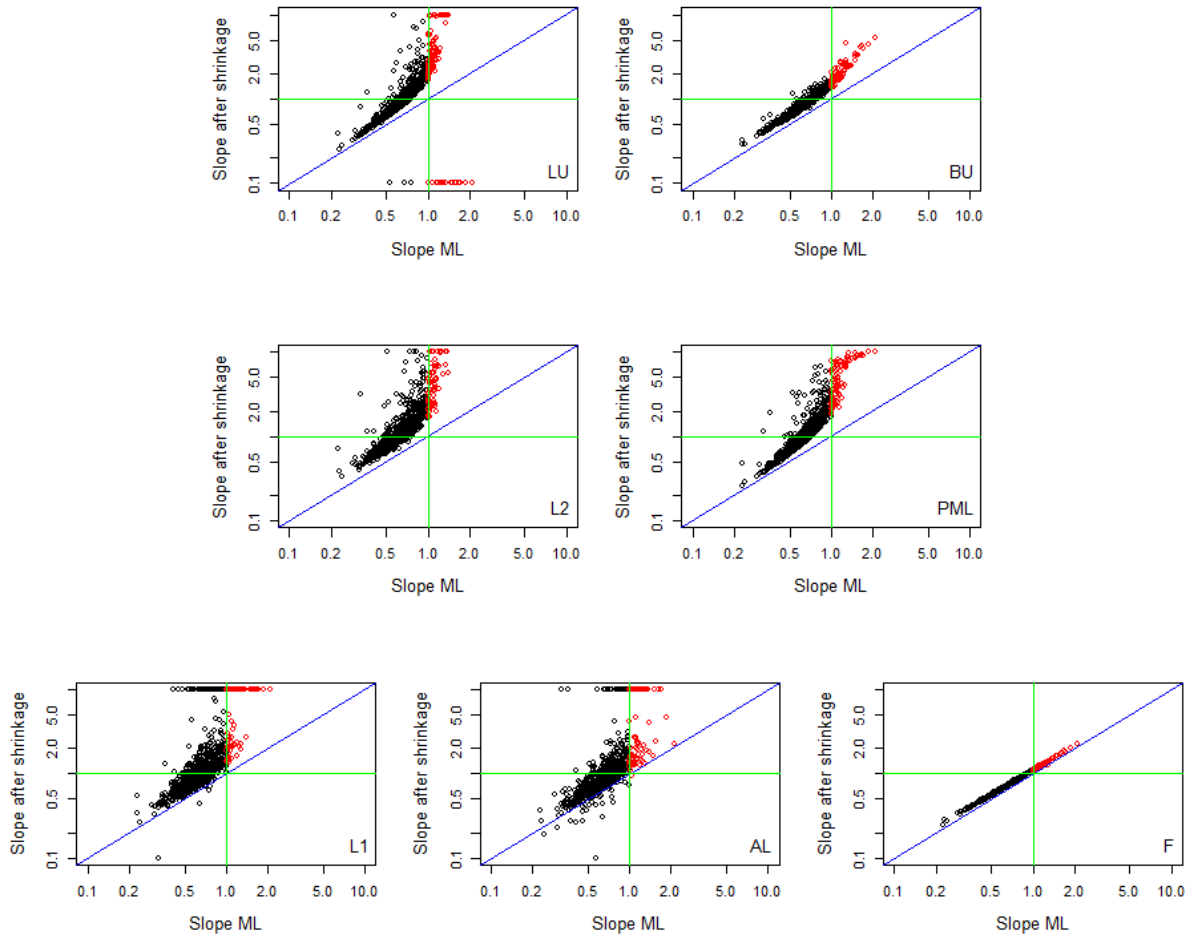
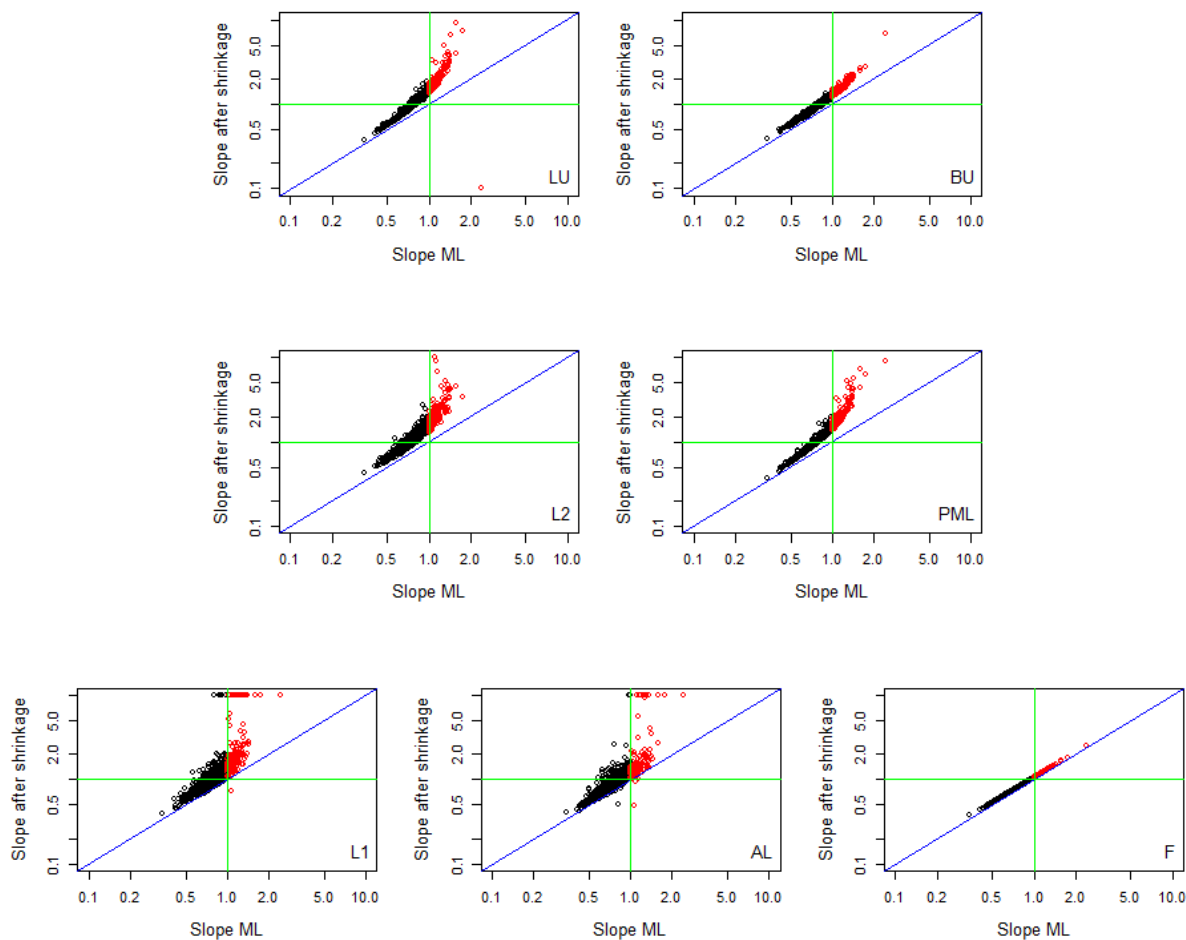


Figure S7. Scatter plots of calibration slopes with shrinkage vs calibration slopes without shrinkage (maximum likelihood). The green lines indicate a slope of 1, which is the target value. The blue line is the diagonal, points on the diagonal had the same calibration slope with and without shrinkage. Red points refer to runs where maximum likelihood gave a slope >1 . ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

I. 5 true predictors, 0 correlation, 10% event rate, 3 events per variable (EPV)

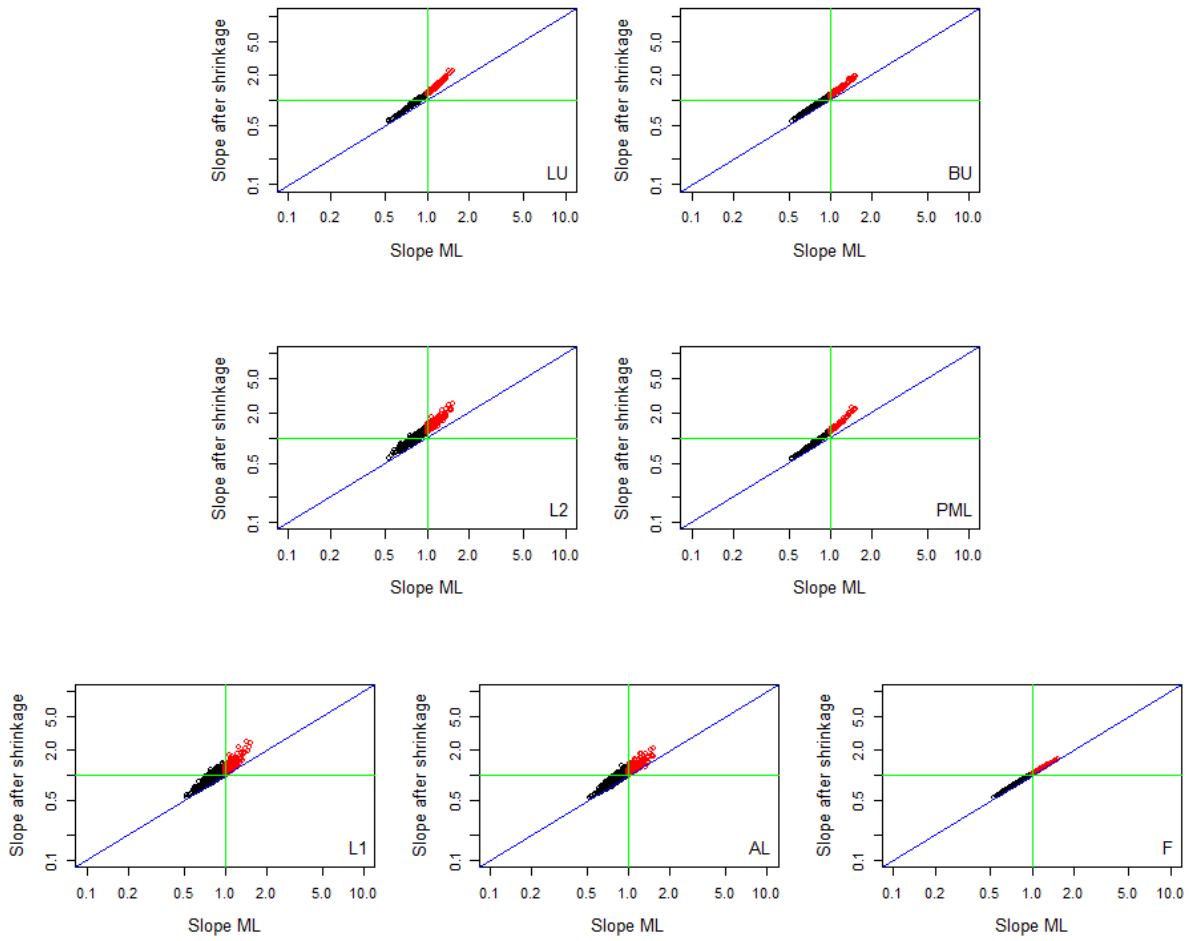


II. 5 true predictors, 0 correlation, 10% event rate, 5 EPV



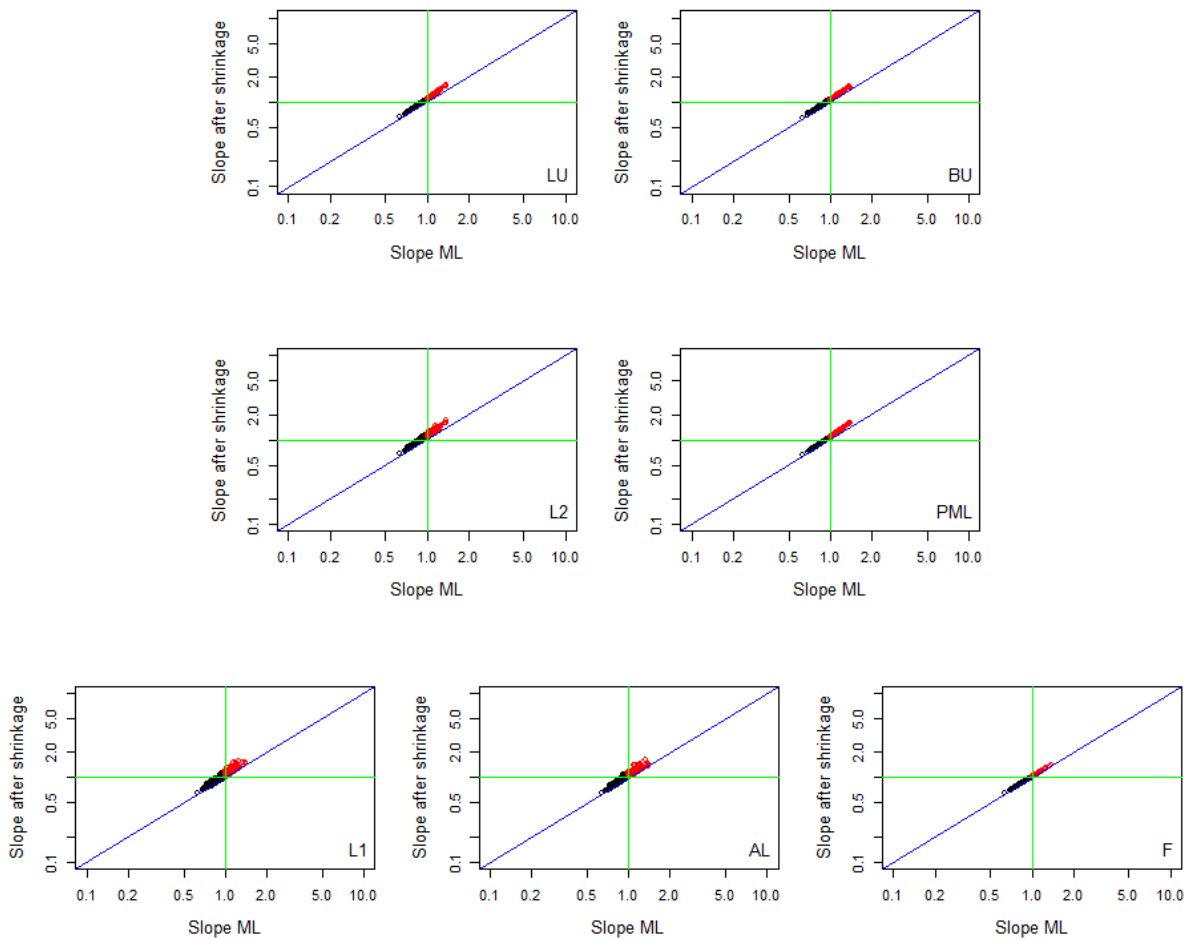
view

III. 5 true predictors, 0 correlation, 10% event rate, 10 EPV



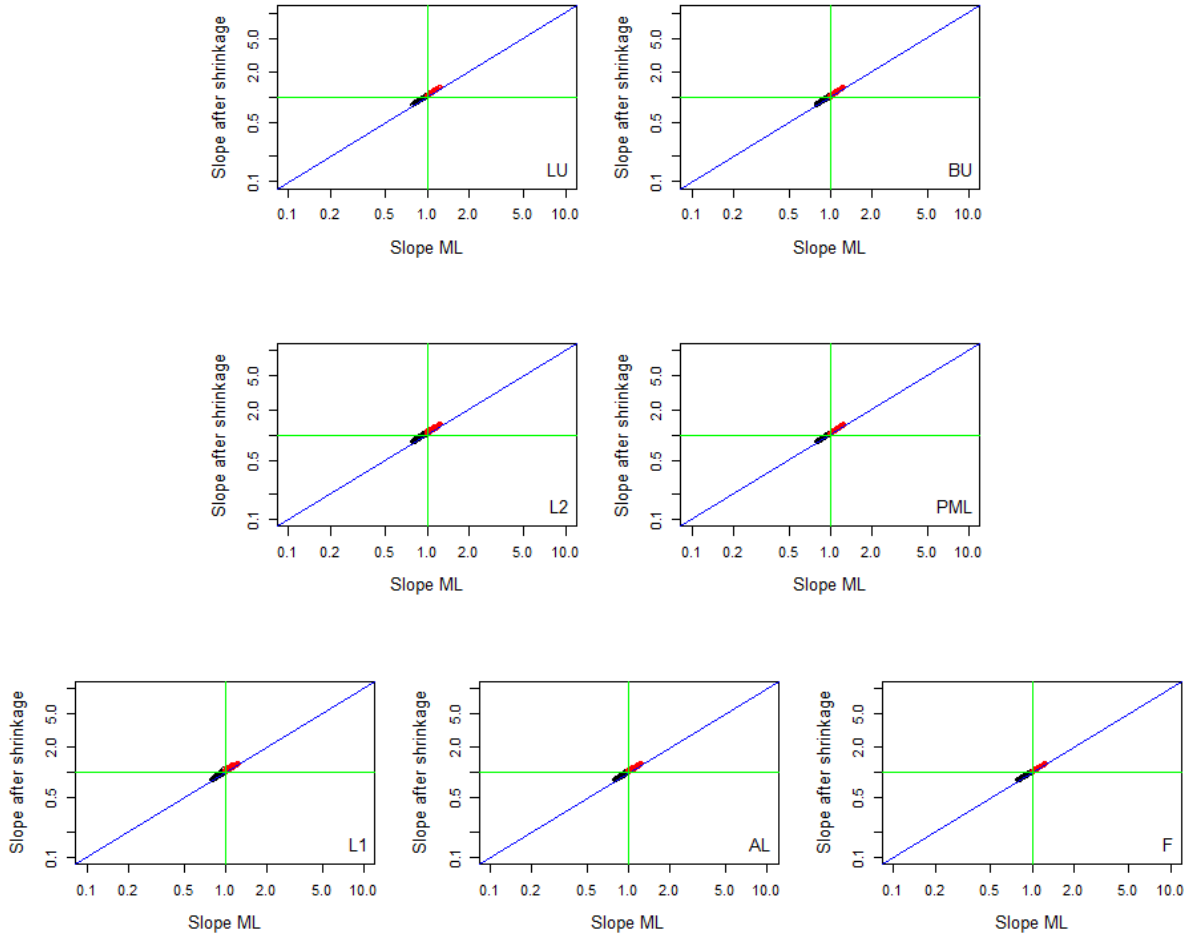
view

IV. 5 true predictors, 0 correlation, 10% event rate, 20 EPV



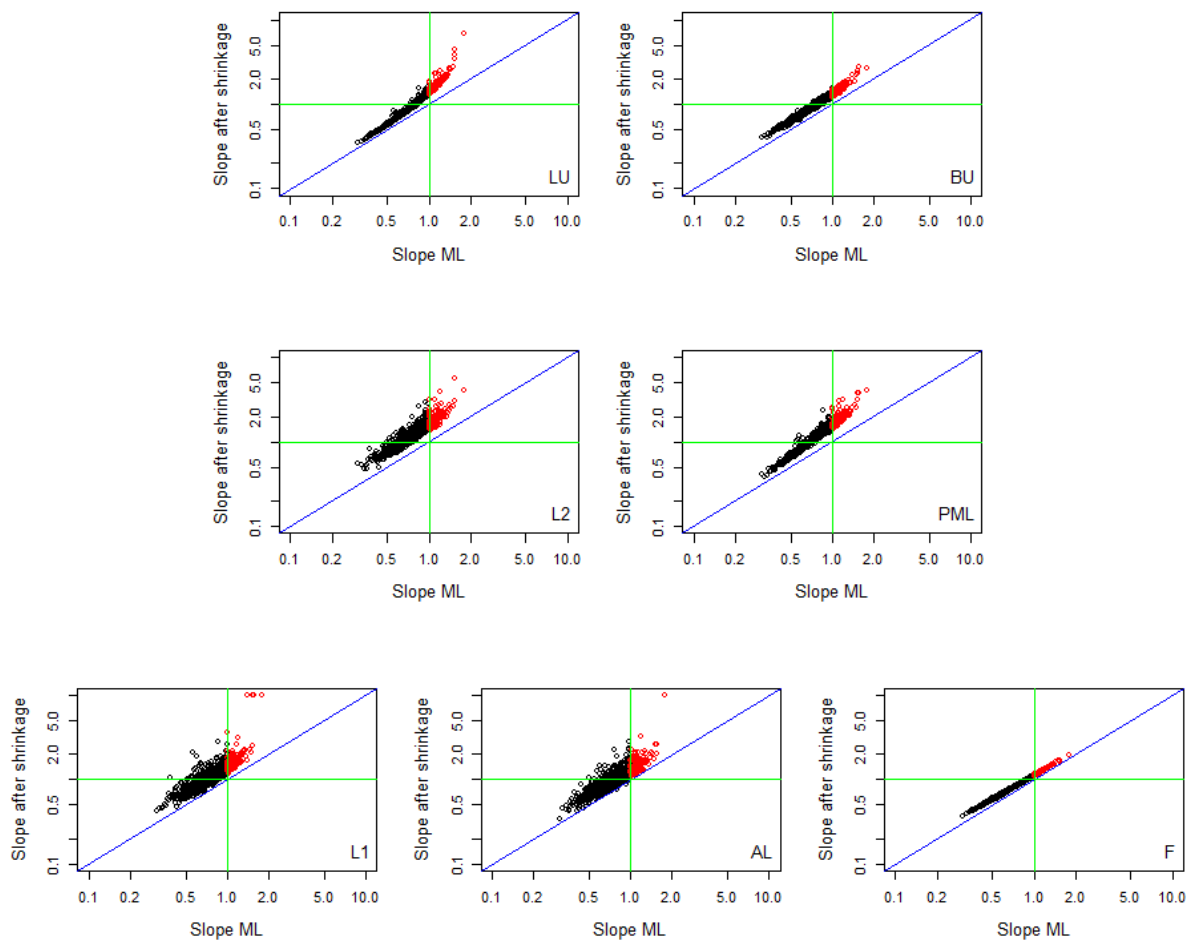
view

V. 5 true predictors, 0 correlation, 10% event rate, 50 EPV



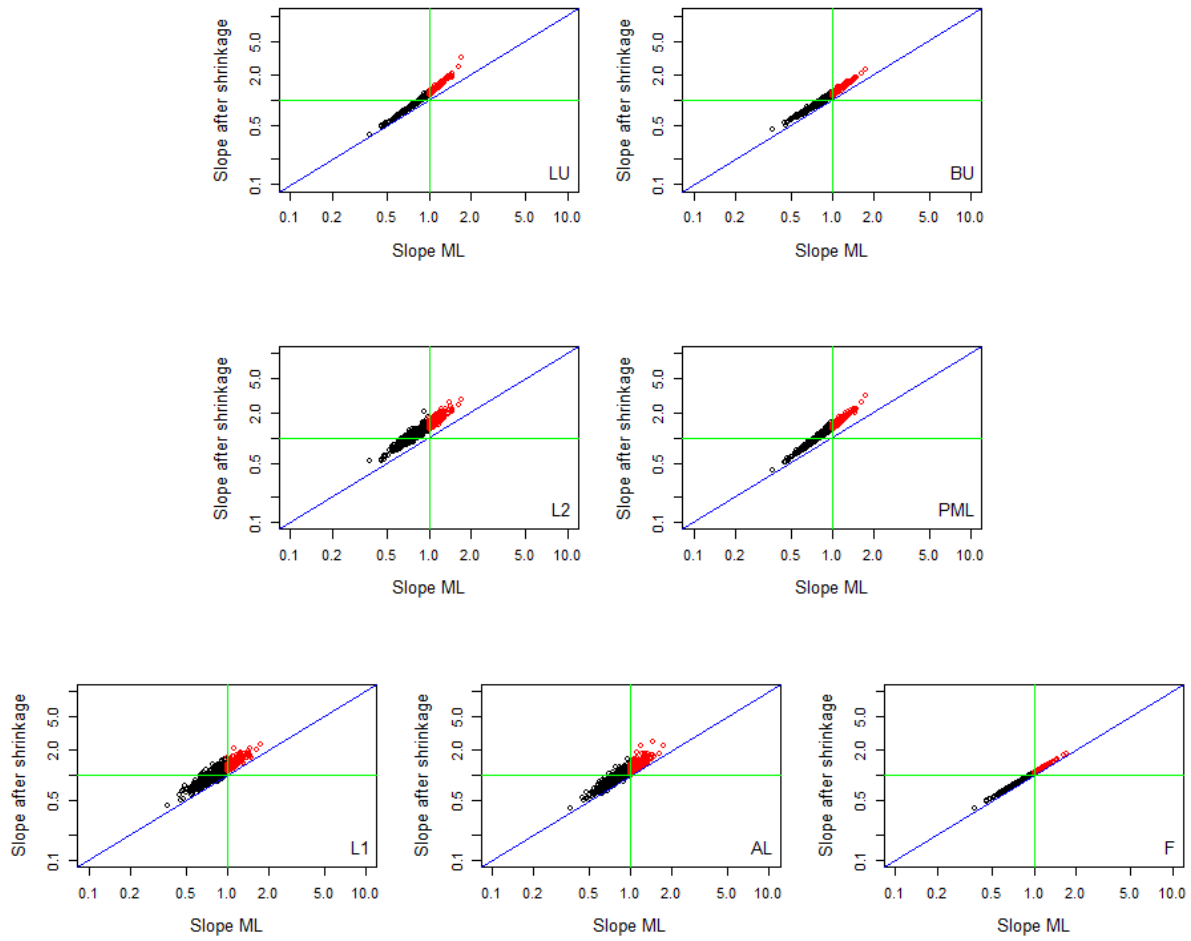
view

VI. 5 true predictors, 0.5 correlation, 10% event rate, 3 EPV



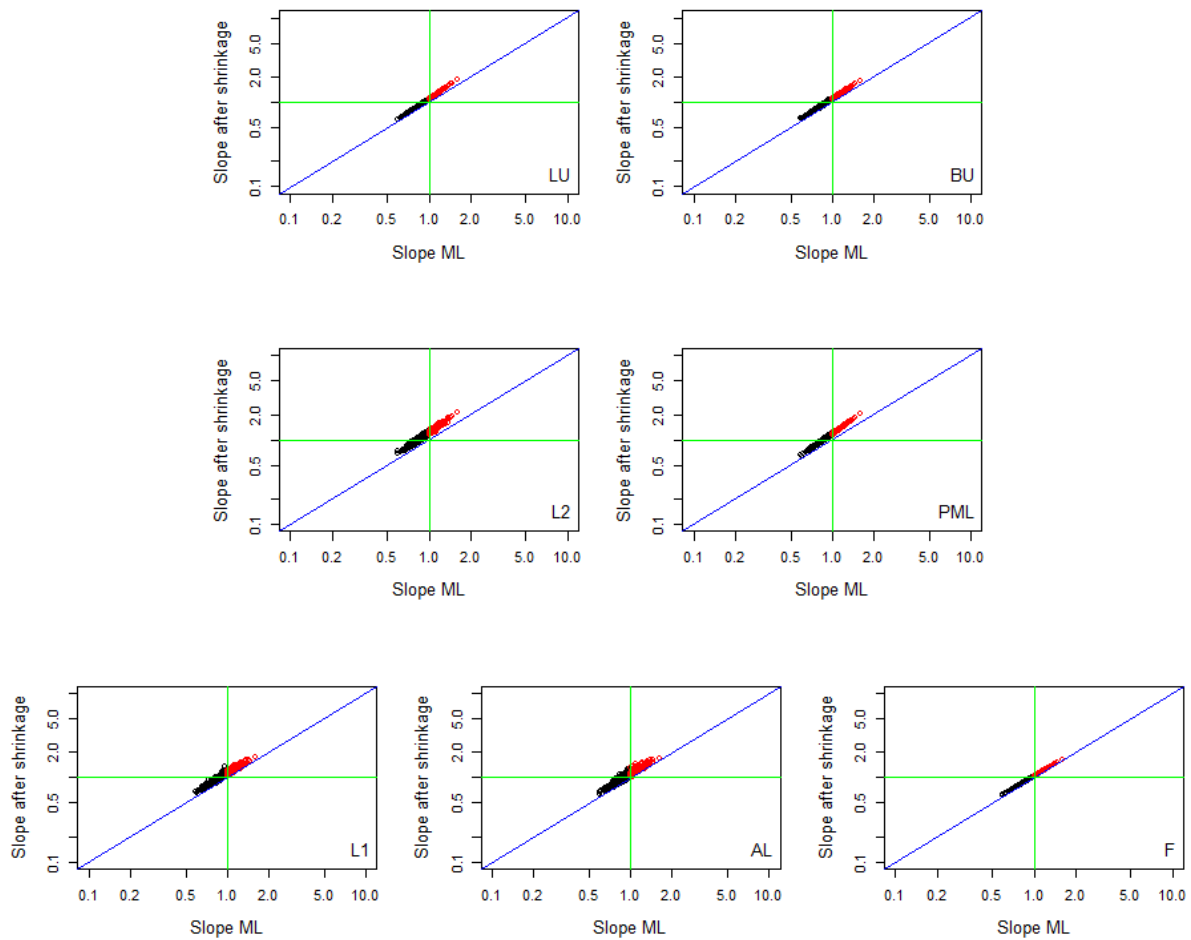
view

VII. 5 true predictors, 0.5 correlation, 10% event rate, 5 EPV



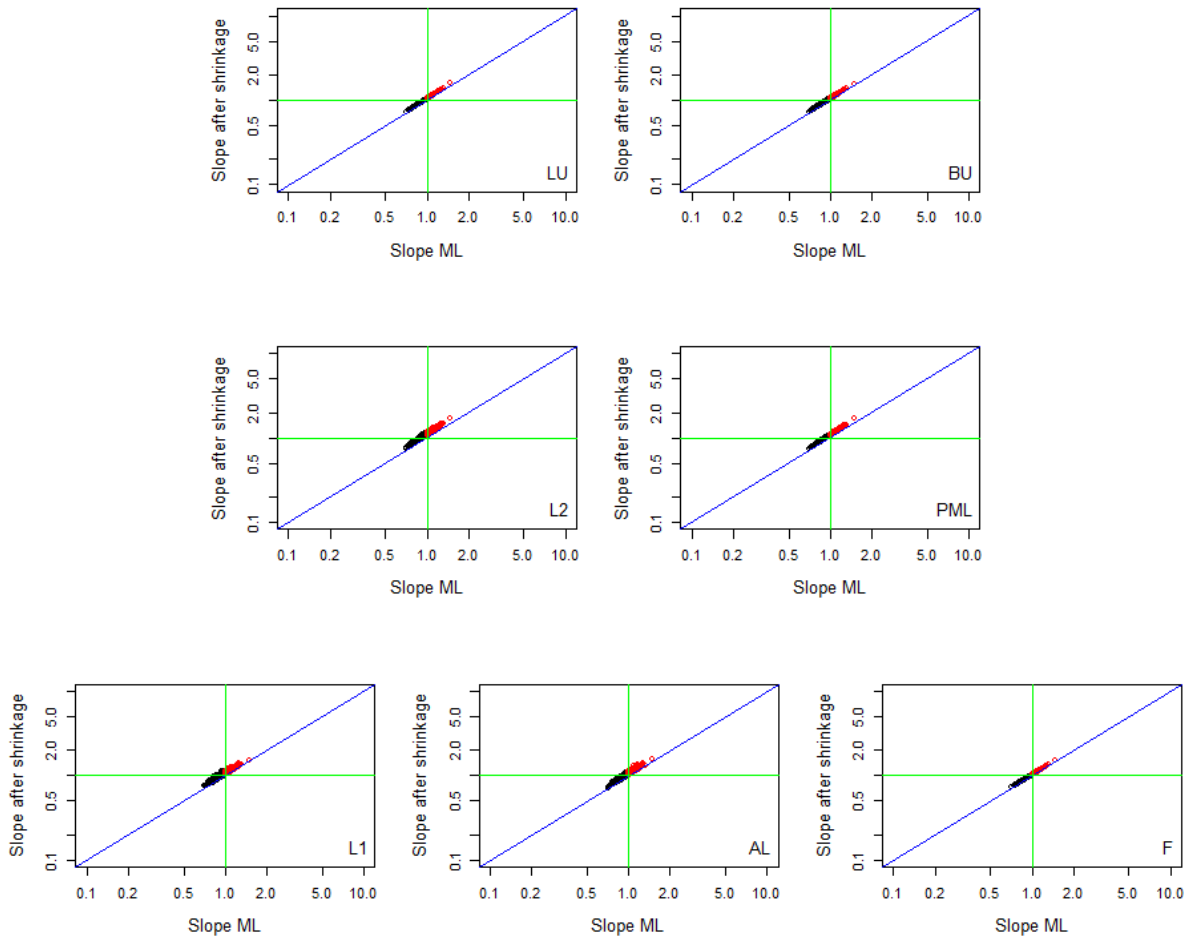
view

VIII. 5 true predictors, 0.5 correlation, 10% event rate, 10 EPV



view

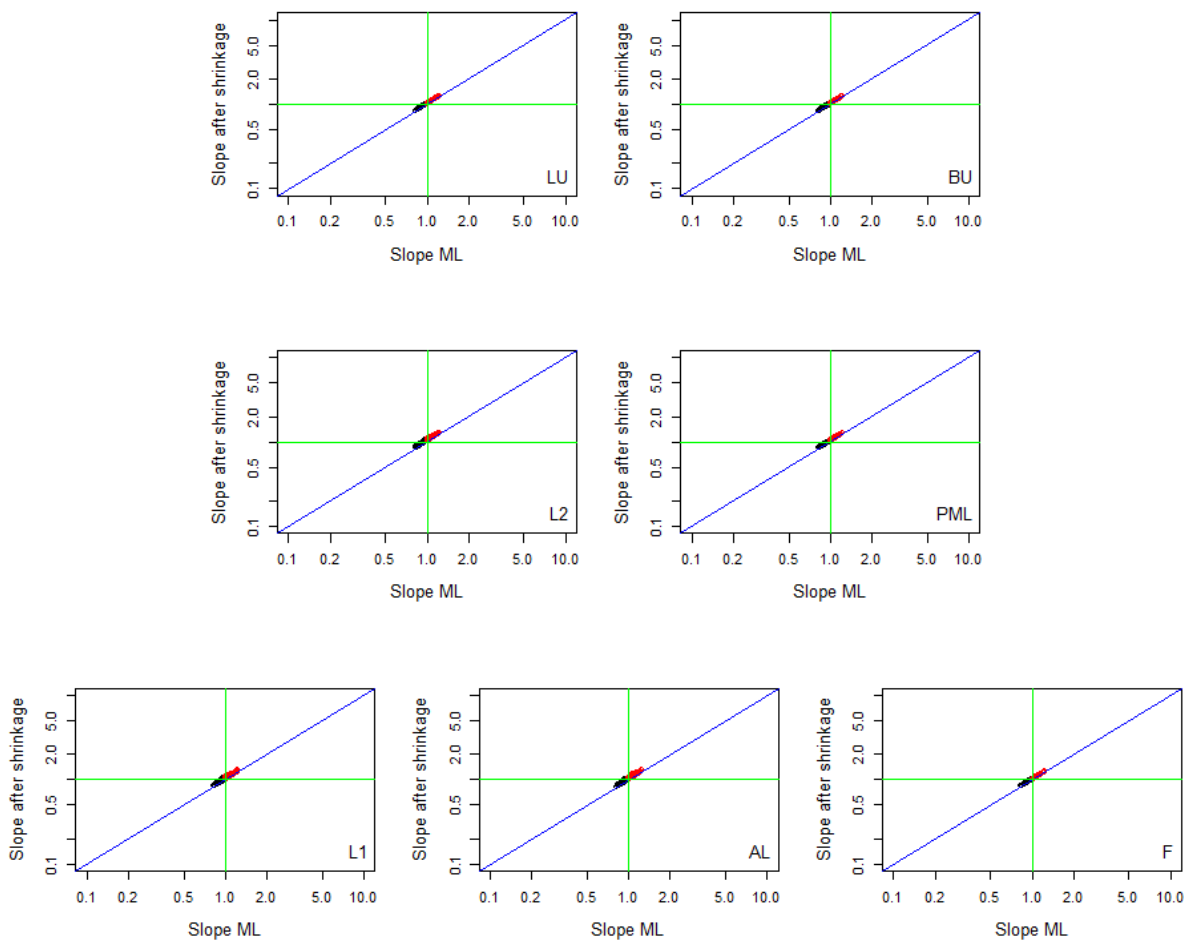
IX. 5 true predictors, 0.5 correlation, 10% event rate, 20 EPV



view

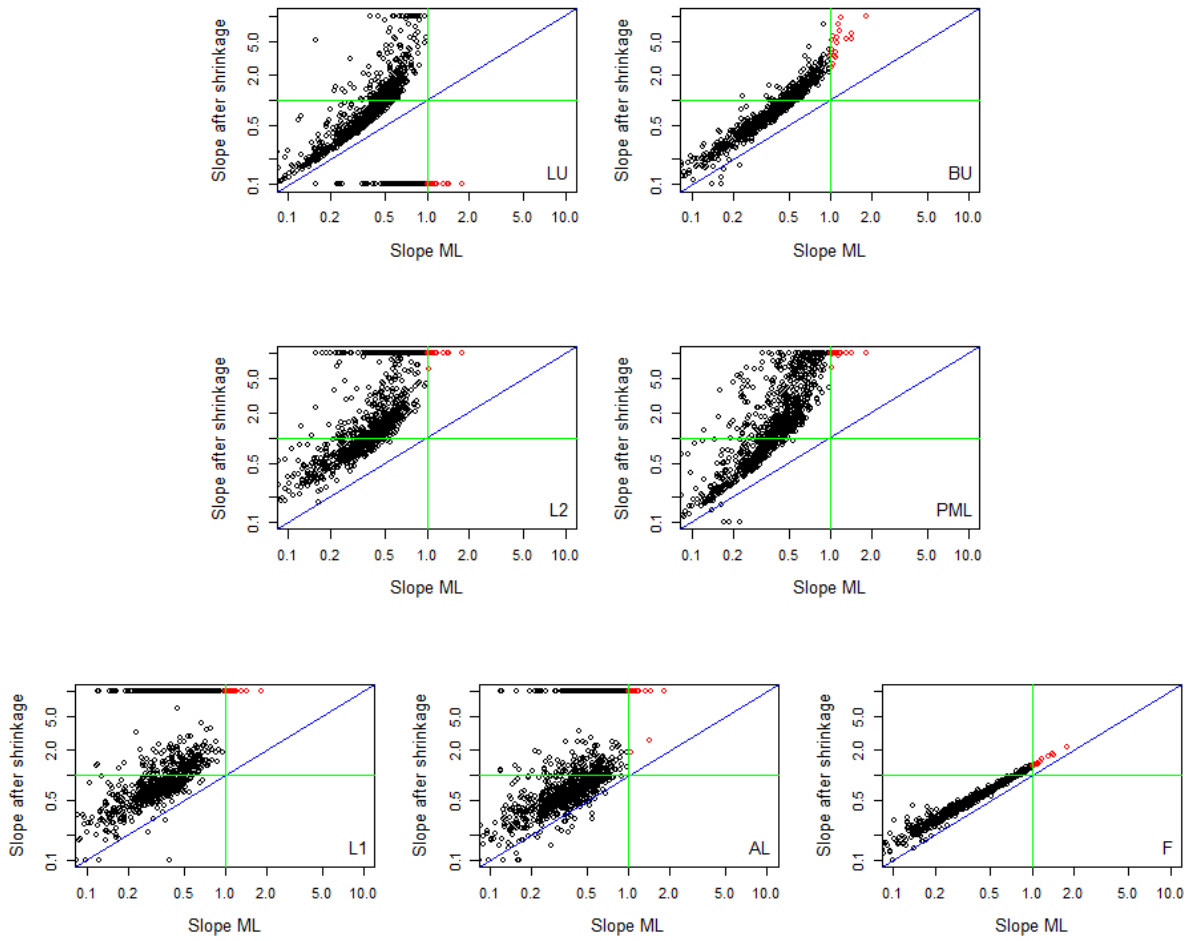
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

X. 5 true predictors, 0.5 correlation, 10% event rate, 50 EPV



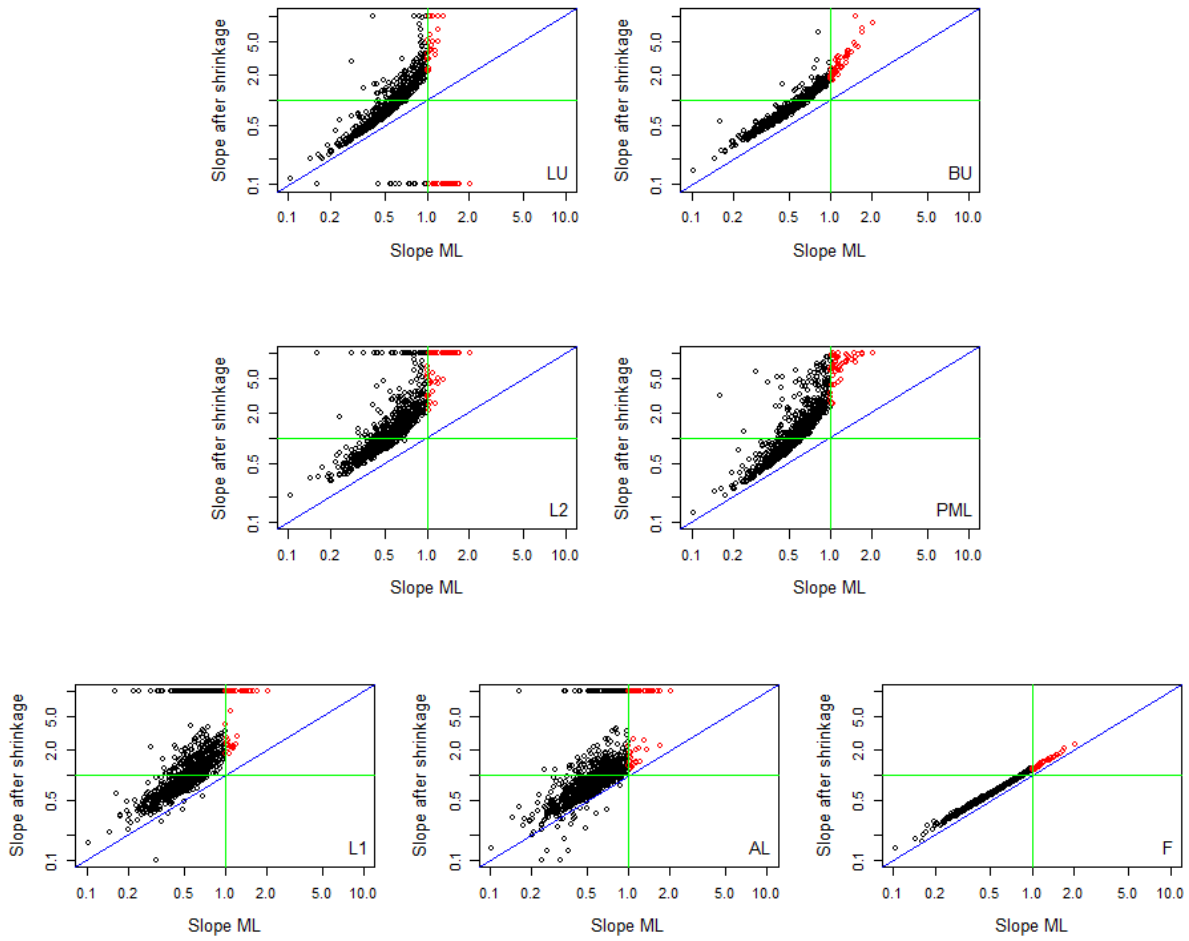
view

XI. 5 true predictors, 0 correlation, 50% event rate, 3 EPV



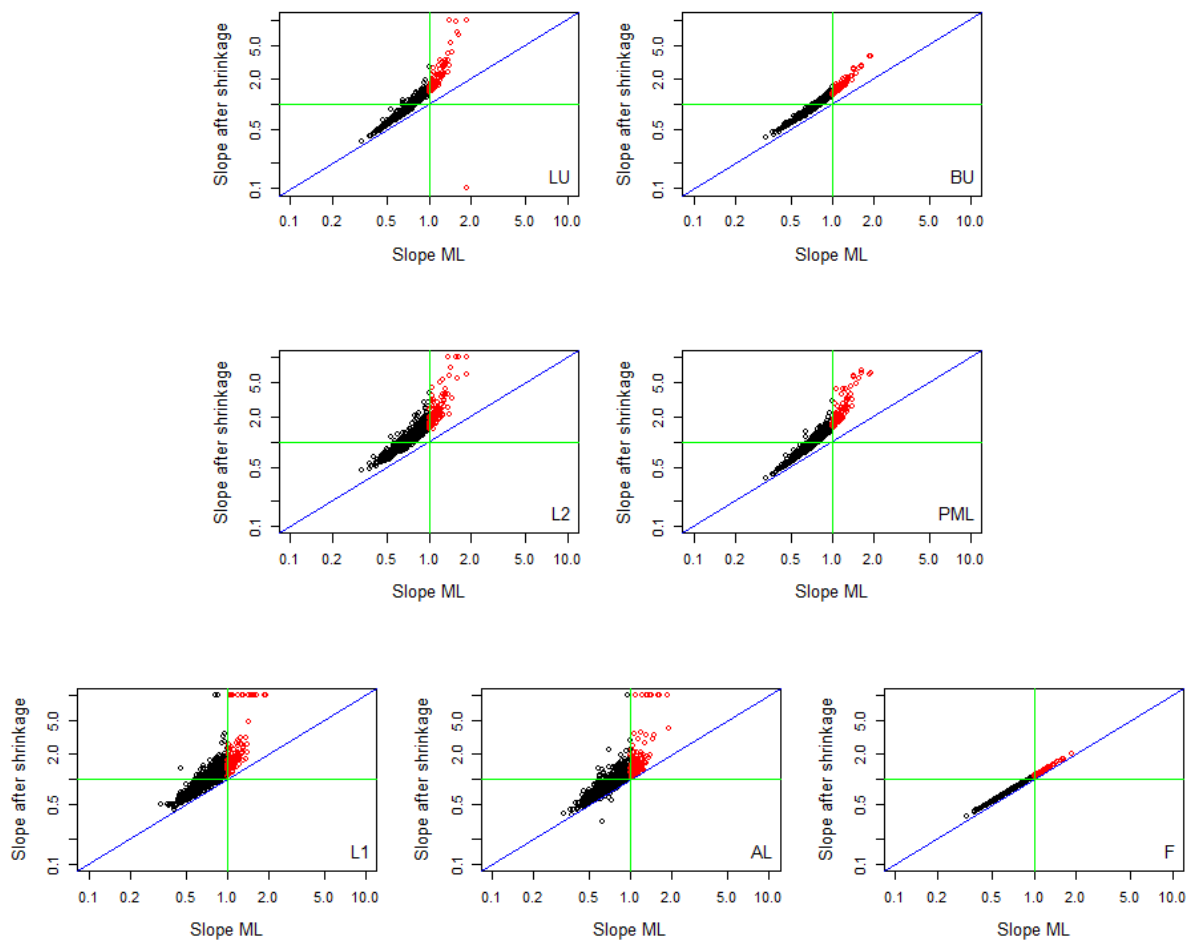
view

XII. 5 true predictors, 0 correlation, 50% event rate, 5 EPV



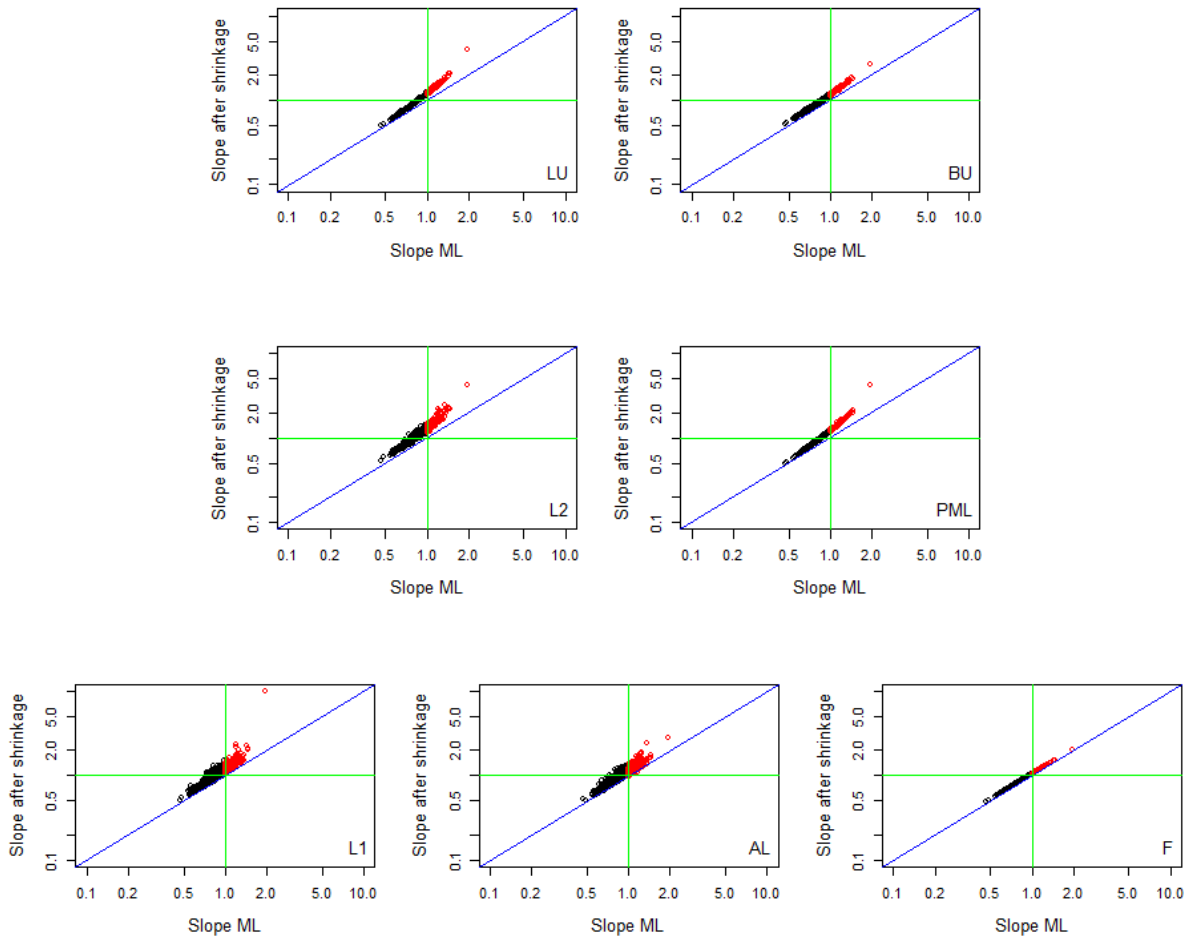
view

XIII. 5 true predictors, 0 correlation, 50% event rate, 10 EPV



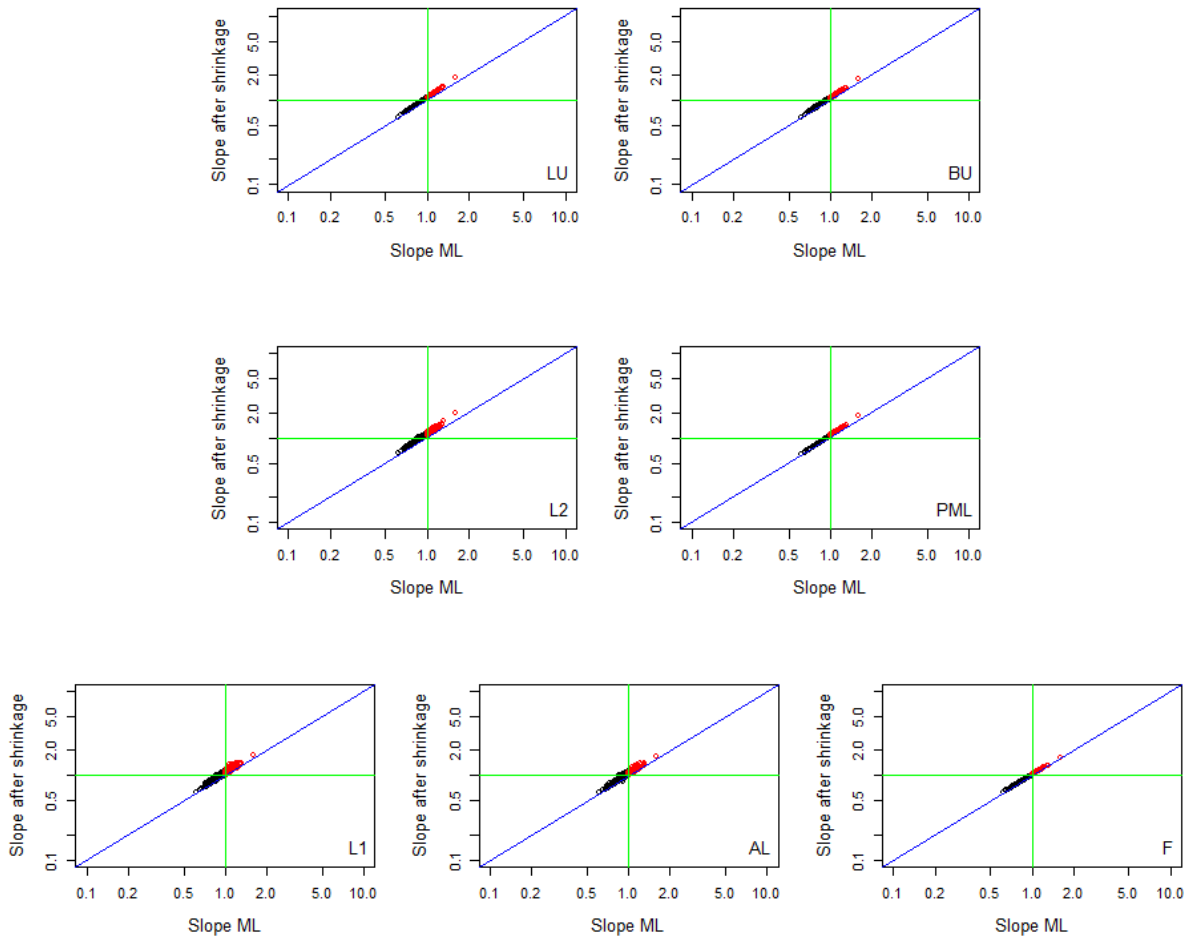
view

XIV. 5 true predictors, 0 correlation, 50% event rate, 20 EPV



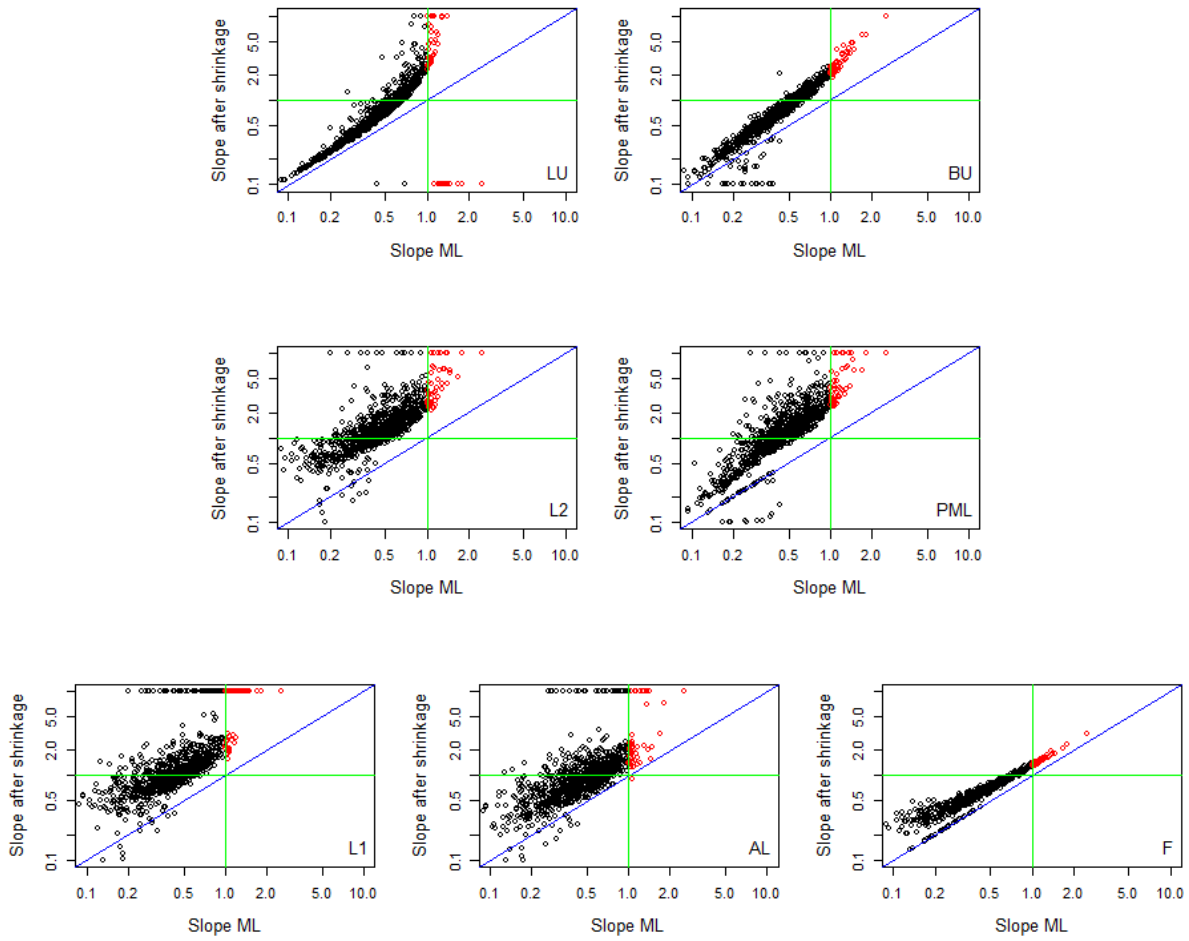
view

XV. 5 true predictors, 0 correlation, 50% event rate, 50 EPV



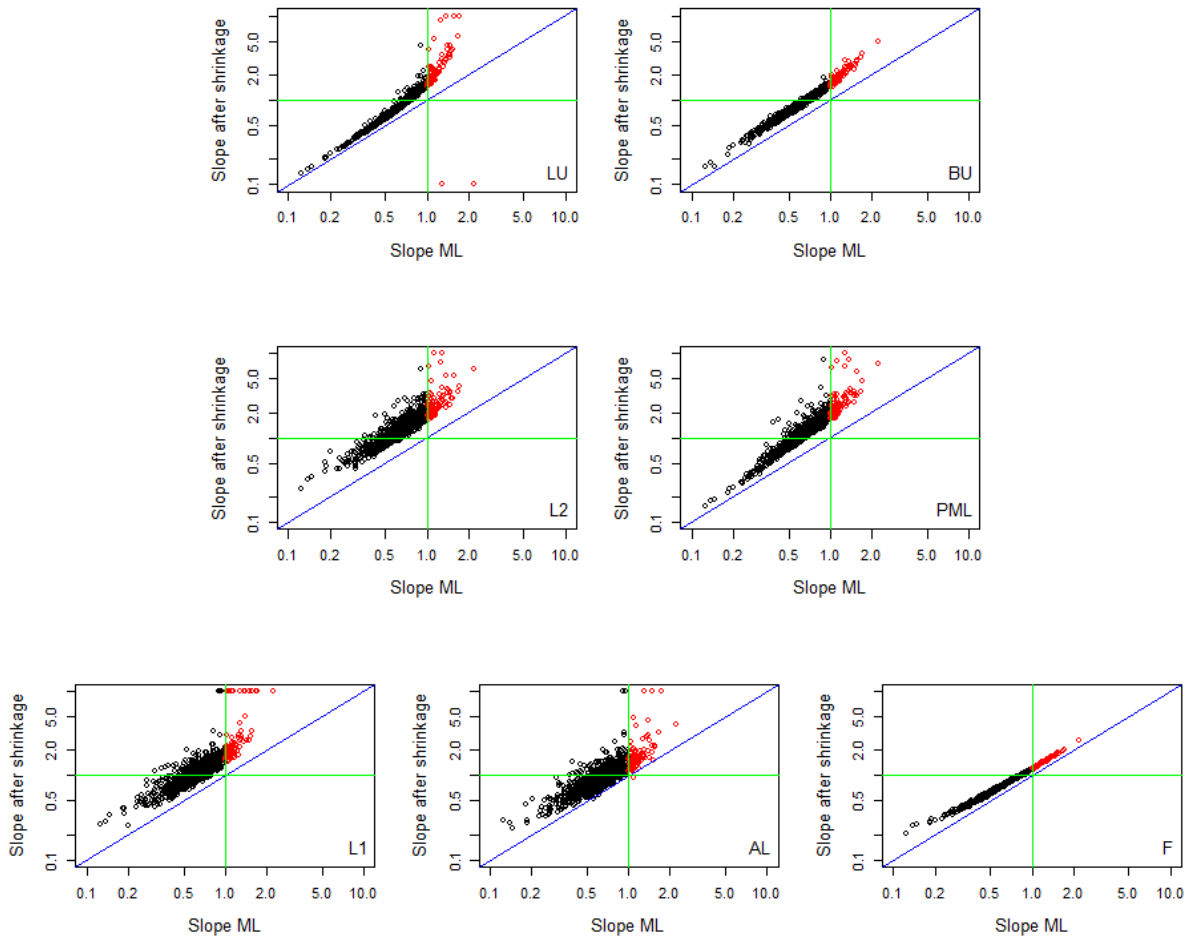
view

XVI. 5 true predictors, 0.5 correlation, 50% event rate, 3 EPV



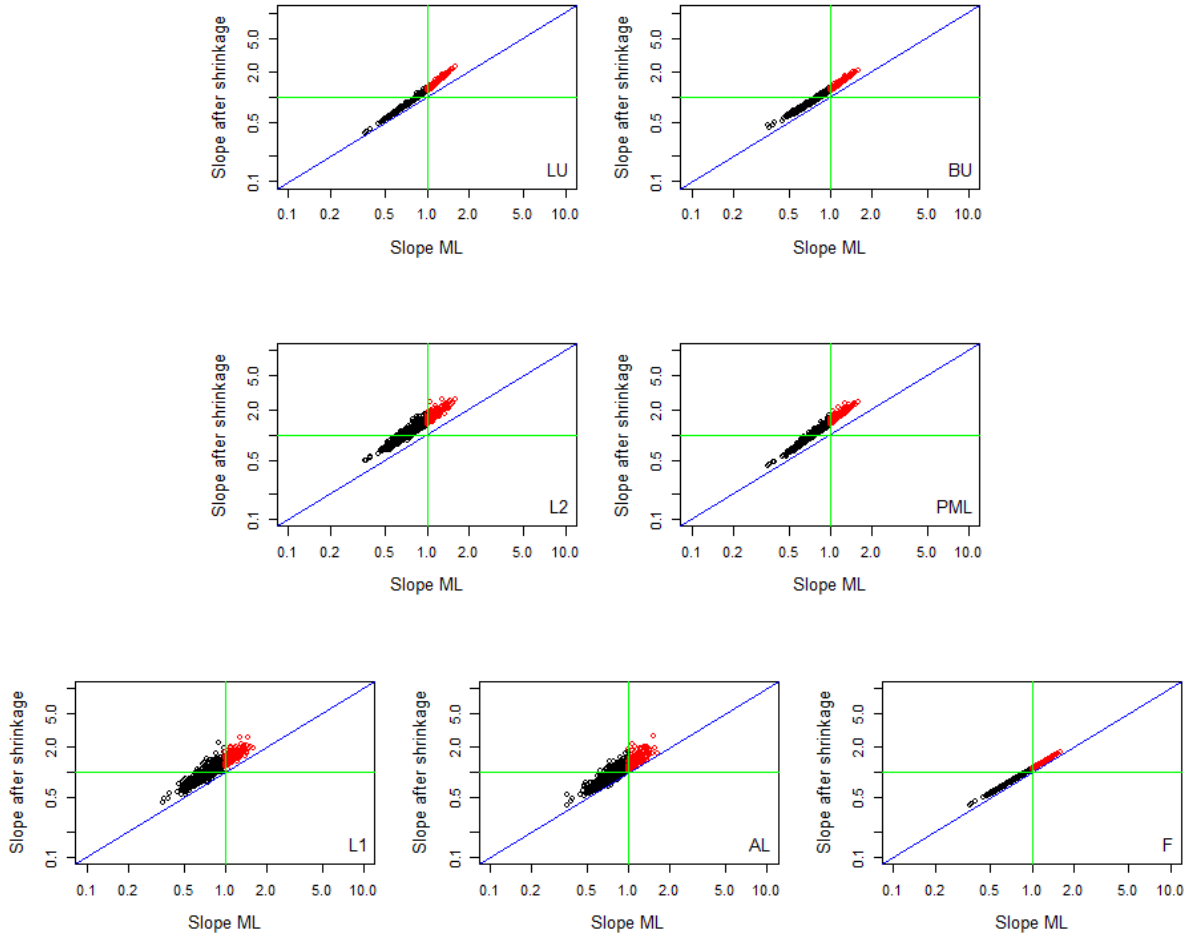
view

XVII. 5 true predictors, 0.5 correlation, 50% event rate, 5 EPV



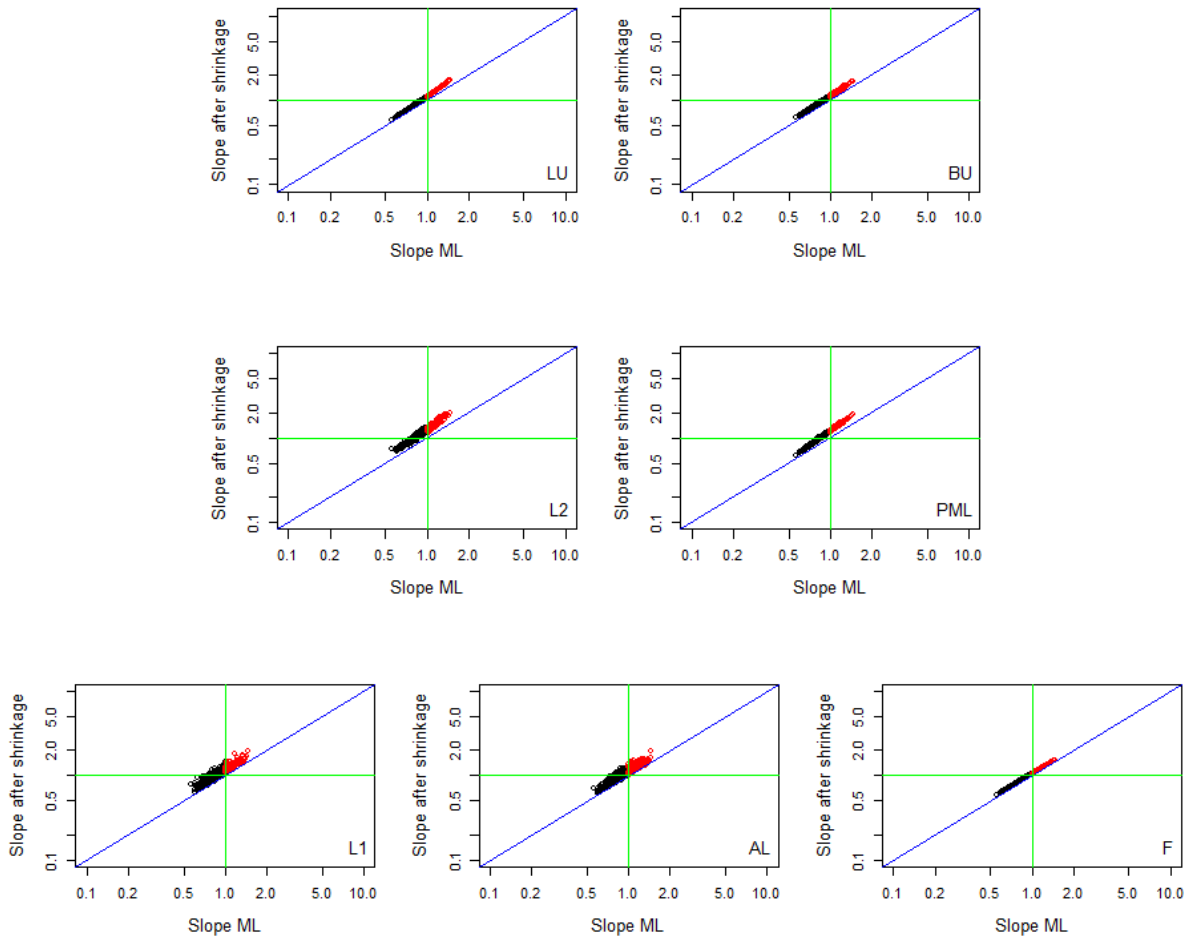
view

XVIII. 5 true predictors, 0.5 correlation, 50% event rate, 10 EPV



view

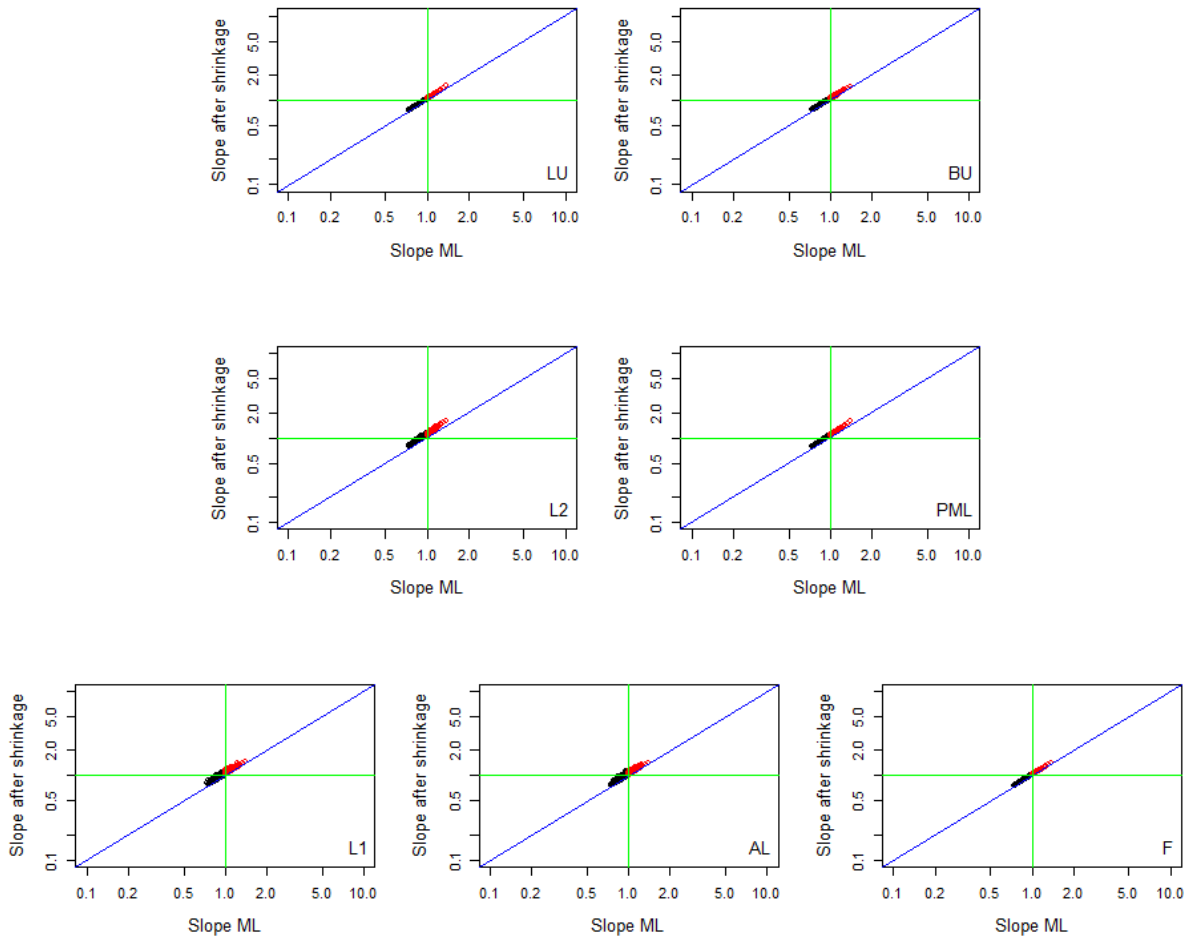
XIX. 5 true predictors, 0.5 correlation, 50% event rate, 20 EPV



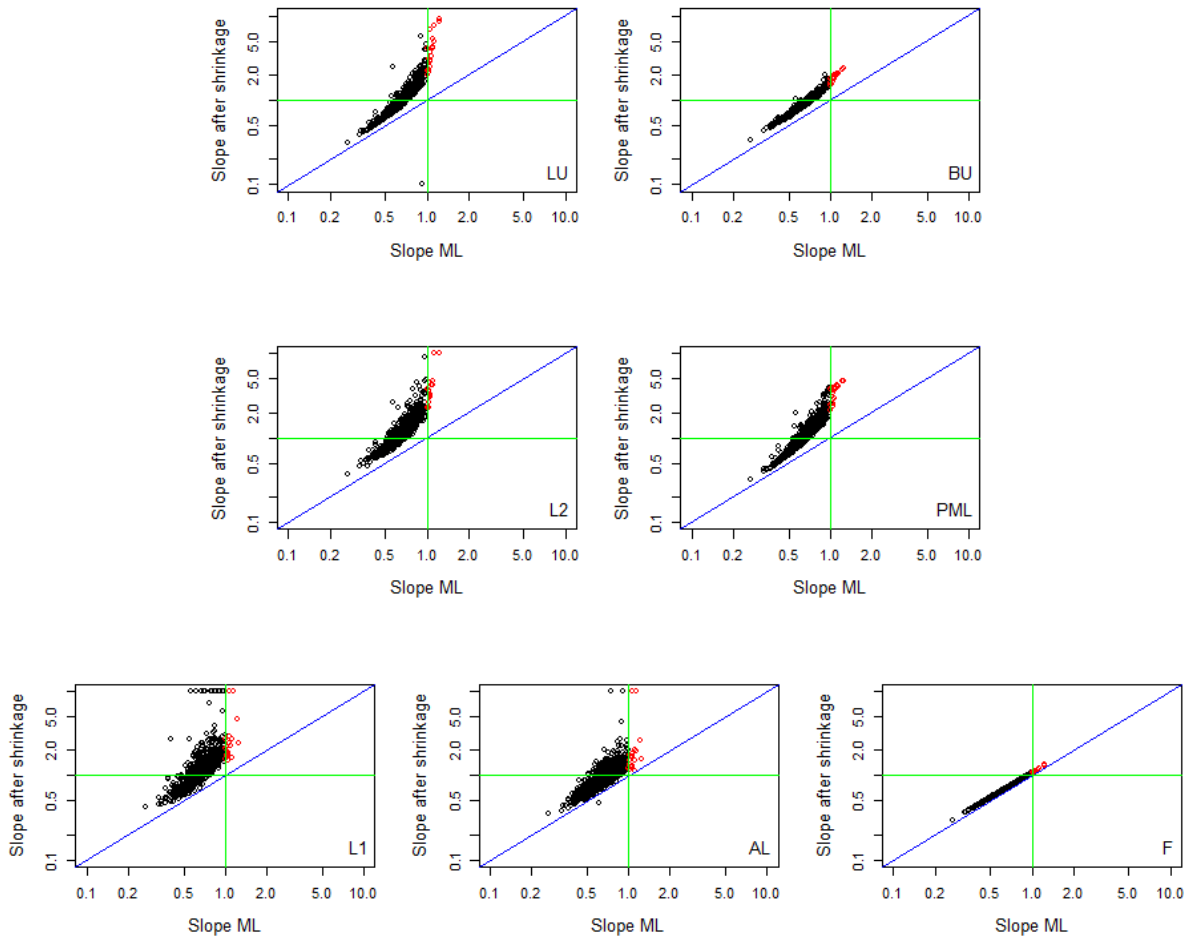
view

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

XX. 5 true predictors, 0.5 correlation, 50% event rate, 50 EPV

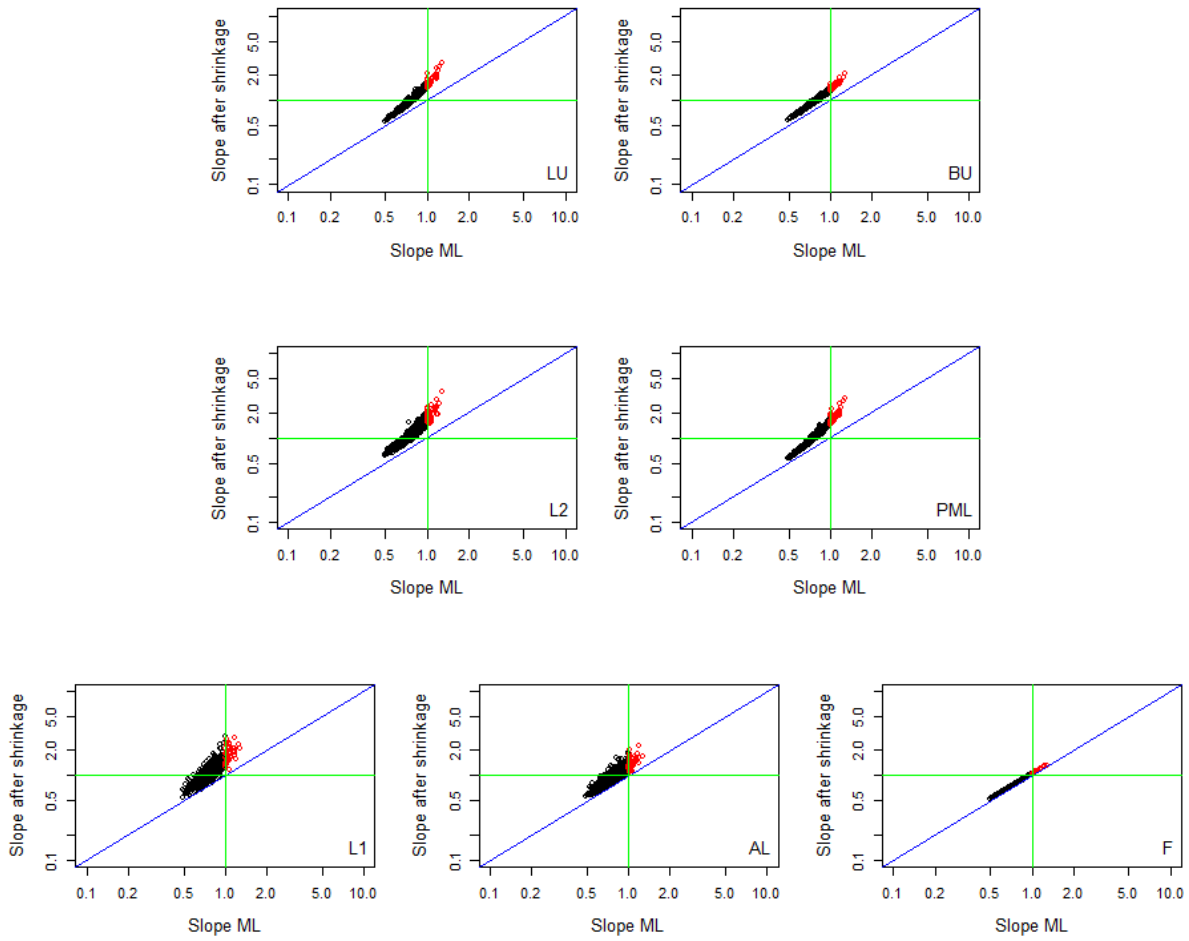


XXI. 5 true and 5 noise predictors, 0 correlation, 10% event rate, 3 EPV



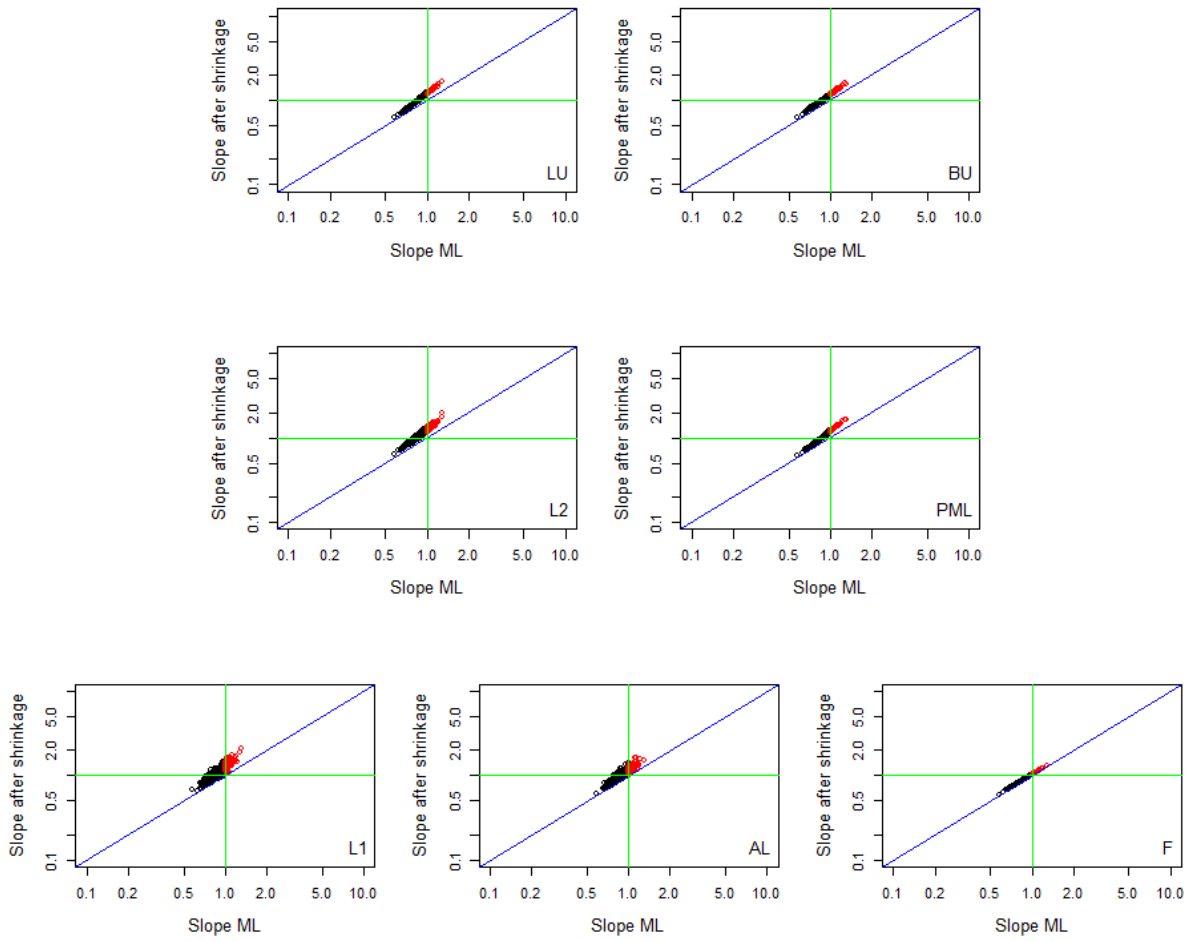
view

XXII. 5 true and 5 noise predictors, 0 correlation, 10% event rate, 5 EPV



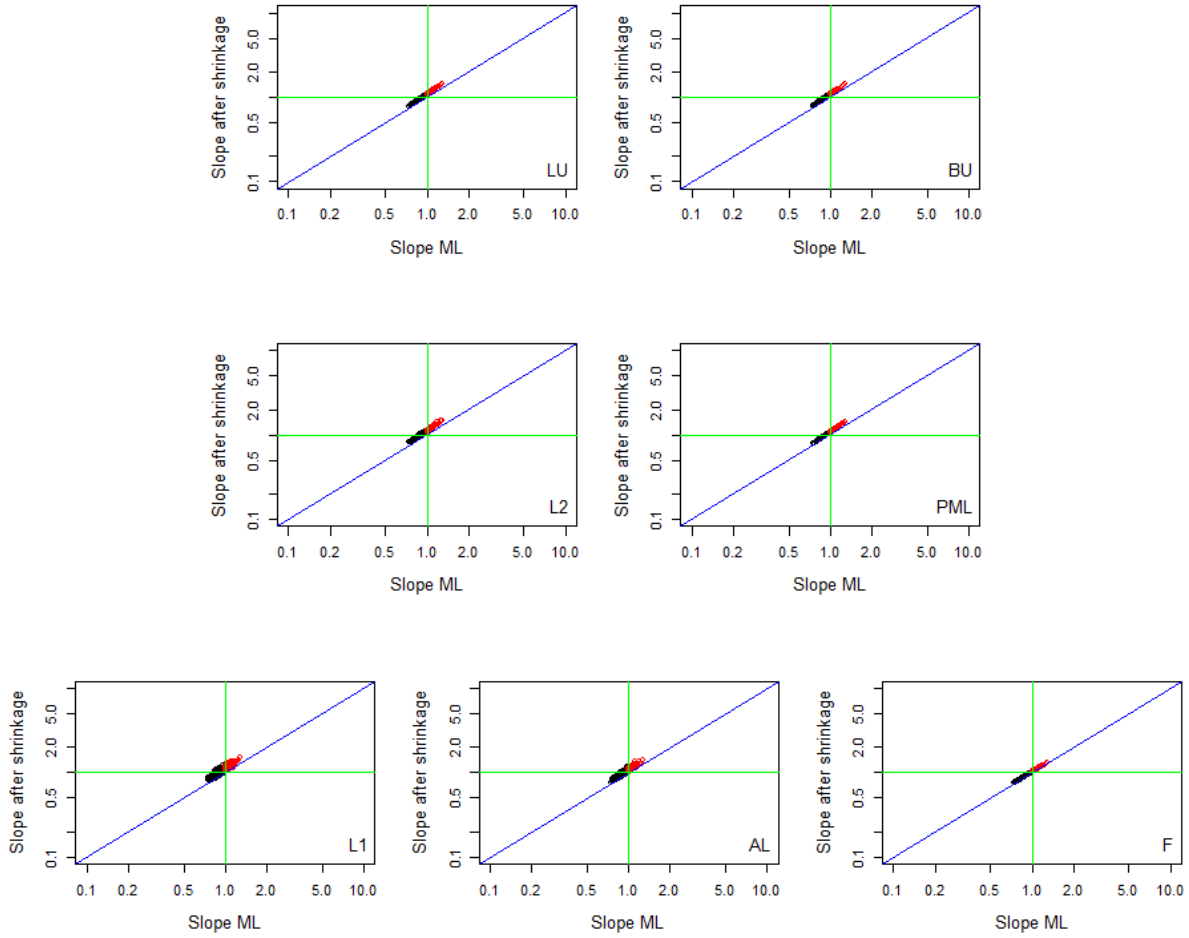
view

XXIII. 5 true and 5 noise predictors, 0 correlation, 10% event rate, 10 EPV



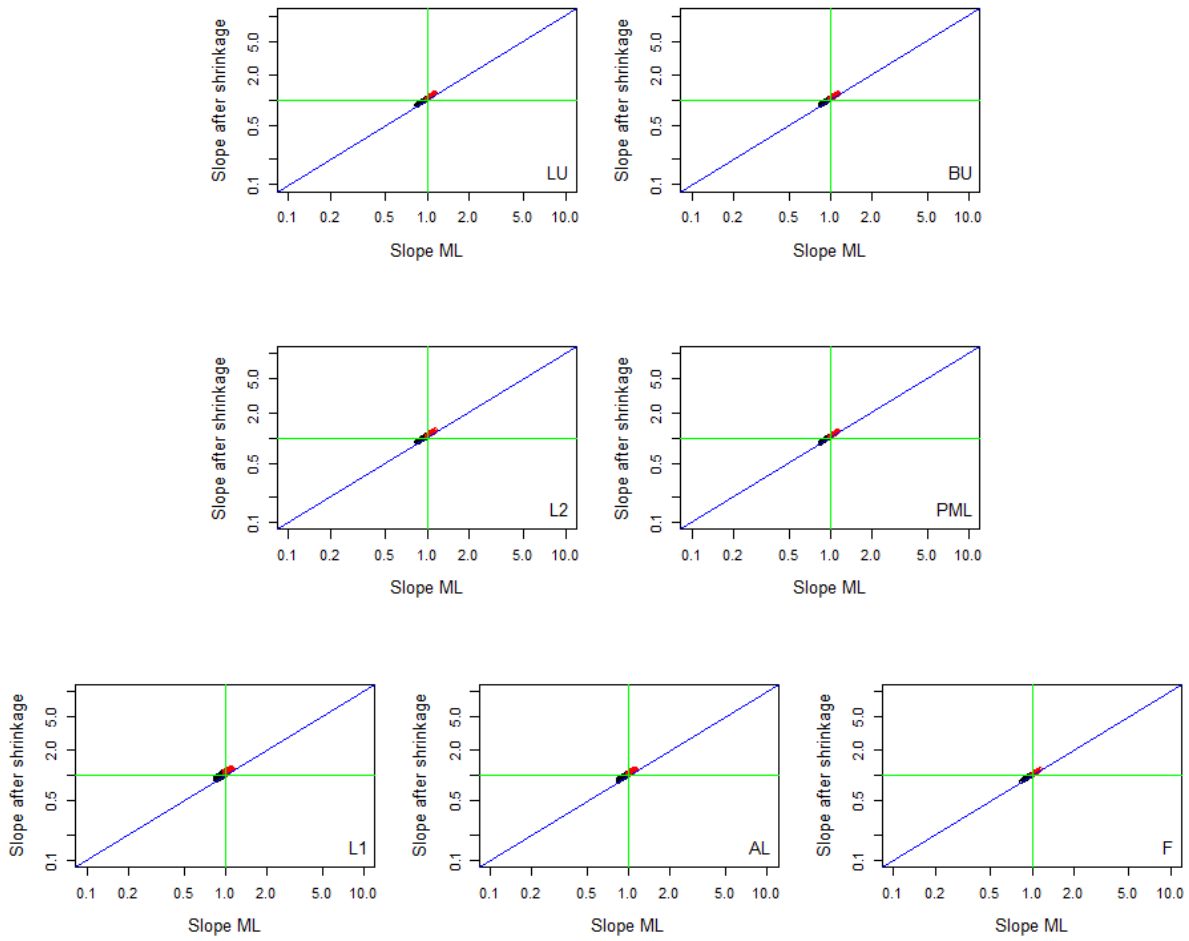
view

XXIV. 5 true and 5 noise predictors, 0 correlation, 10% event rate, 20 EPV



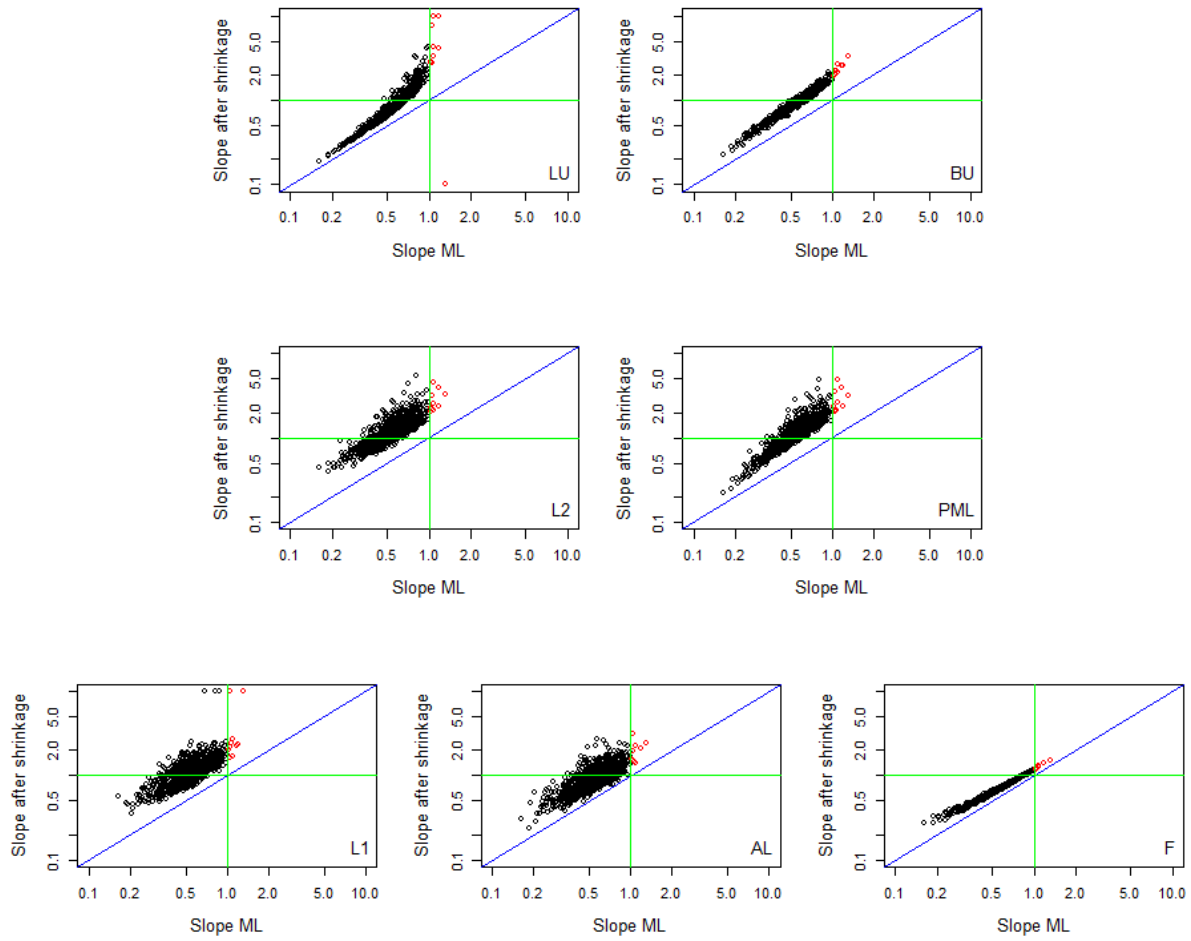
view

XXV. 5 true and 5 noise predictors, 0 correlation, 10% event rate, 50 EPV



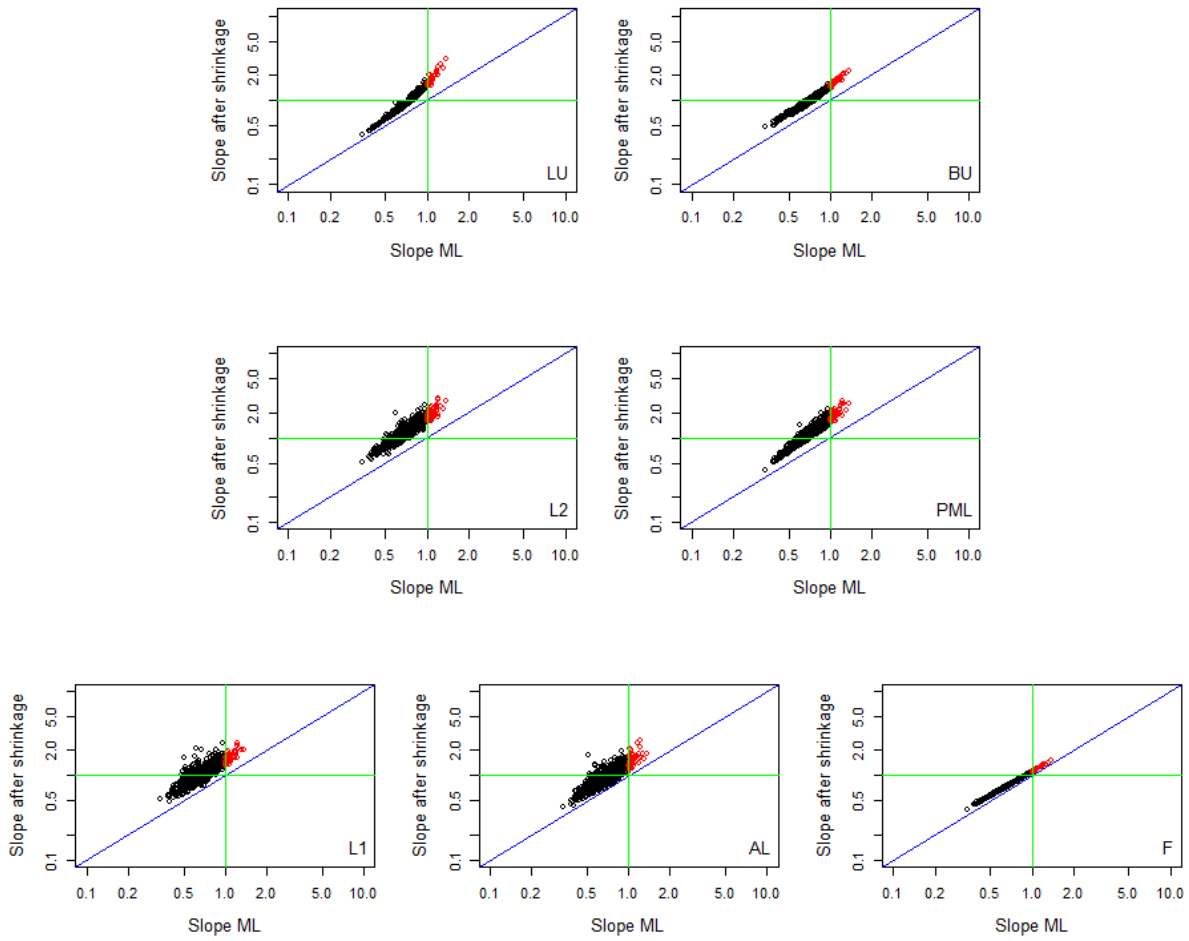
view

XXVI. 5 true and 5 noise predictors, 0.5 correlation, 10% event rate, 3 EPV



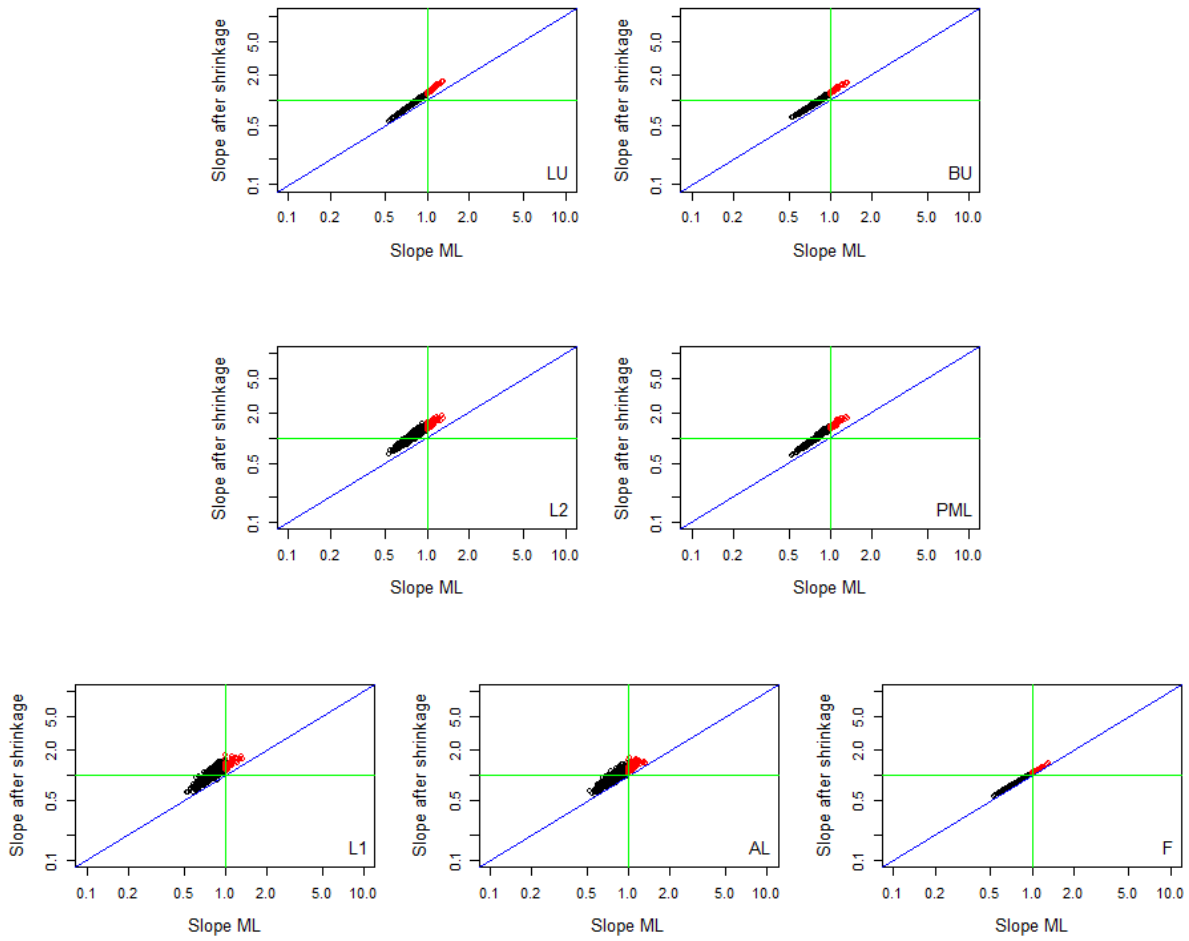
view

XXVII. 5 true and 5 noise predictors, 0.5 correlation, 10% event rate, 5 EPV



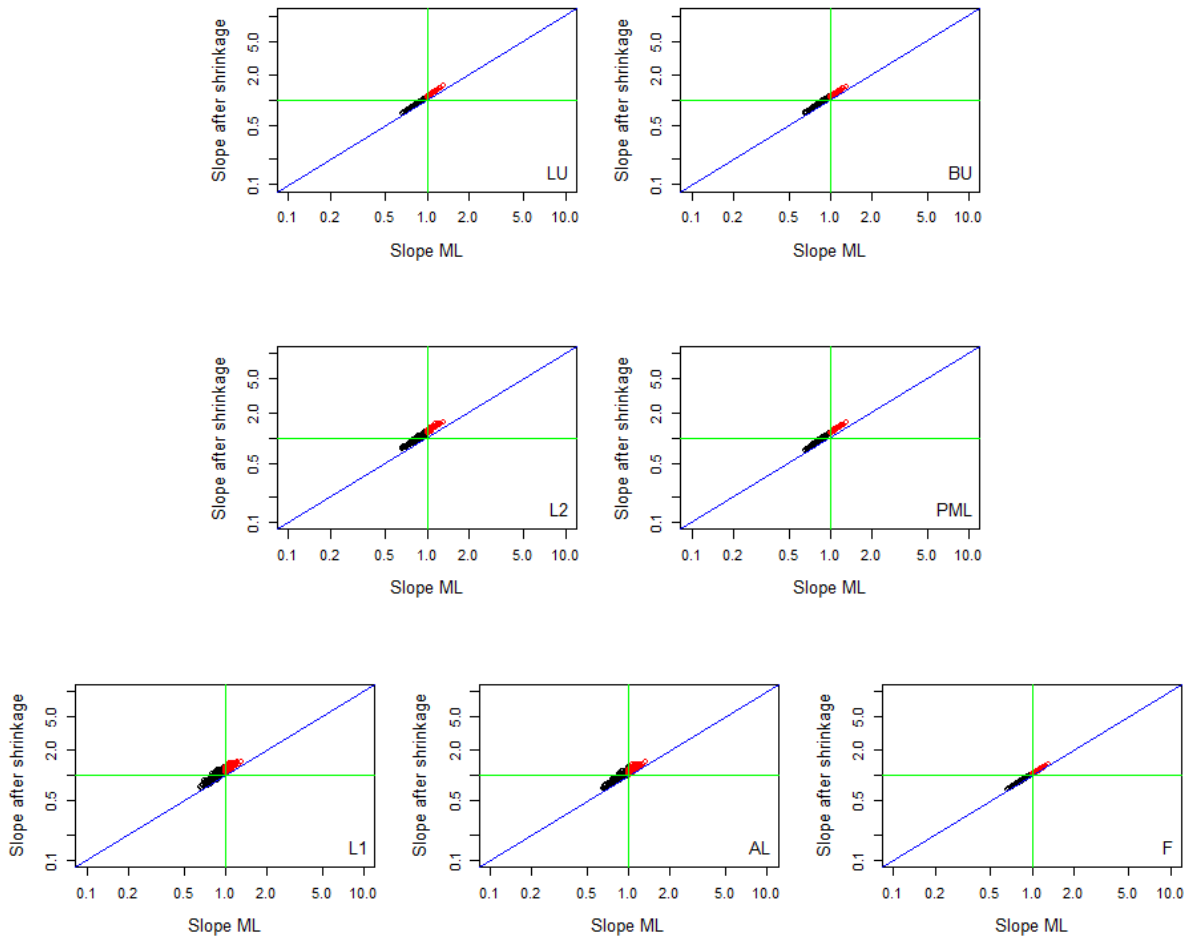
view

XXVIII. 5 true and 5 noise predictors, 0.5 correlation, 10% event rate, 10 EPV



view

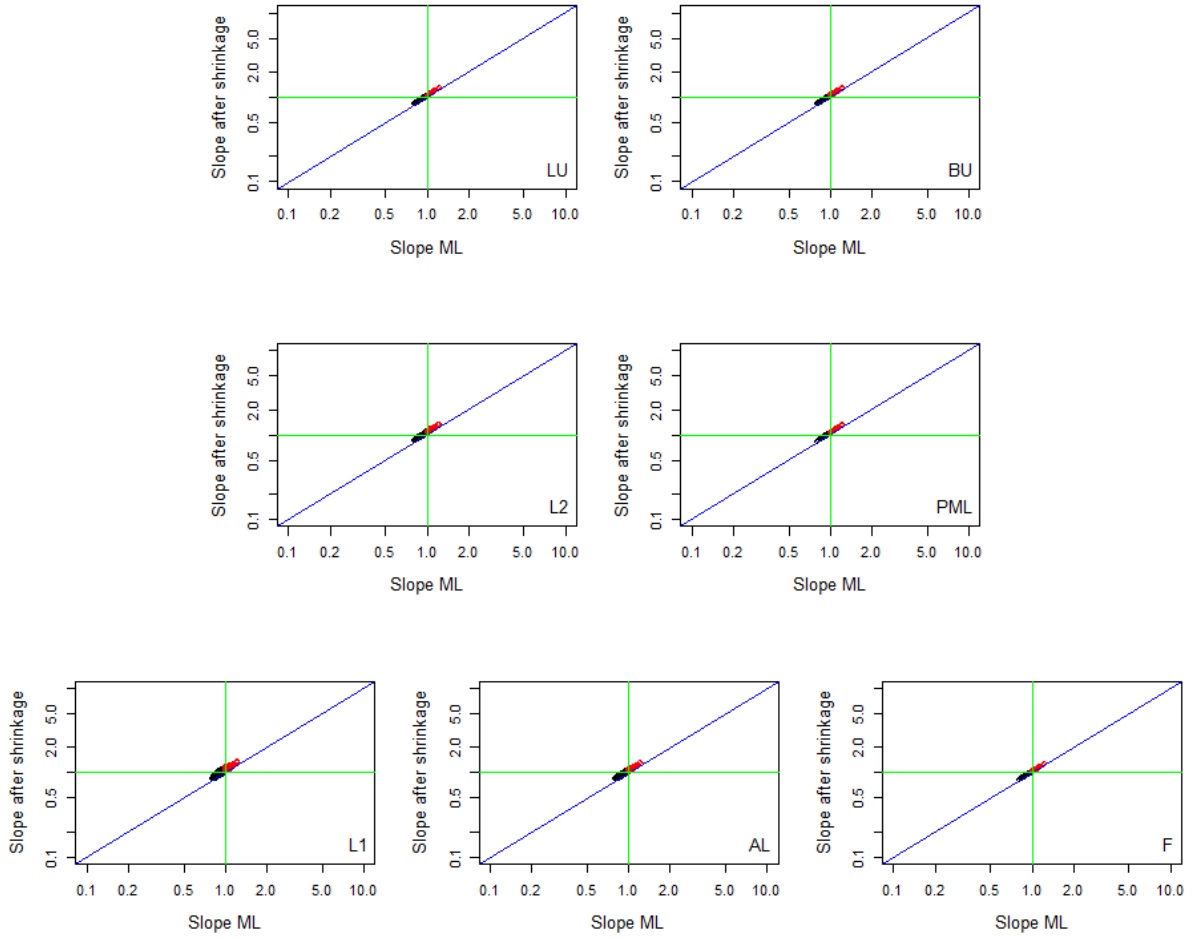
XXIX. 5 true and 5 noise predictors, 0.5 correlation, 10% event rate, 20 EPV



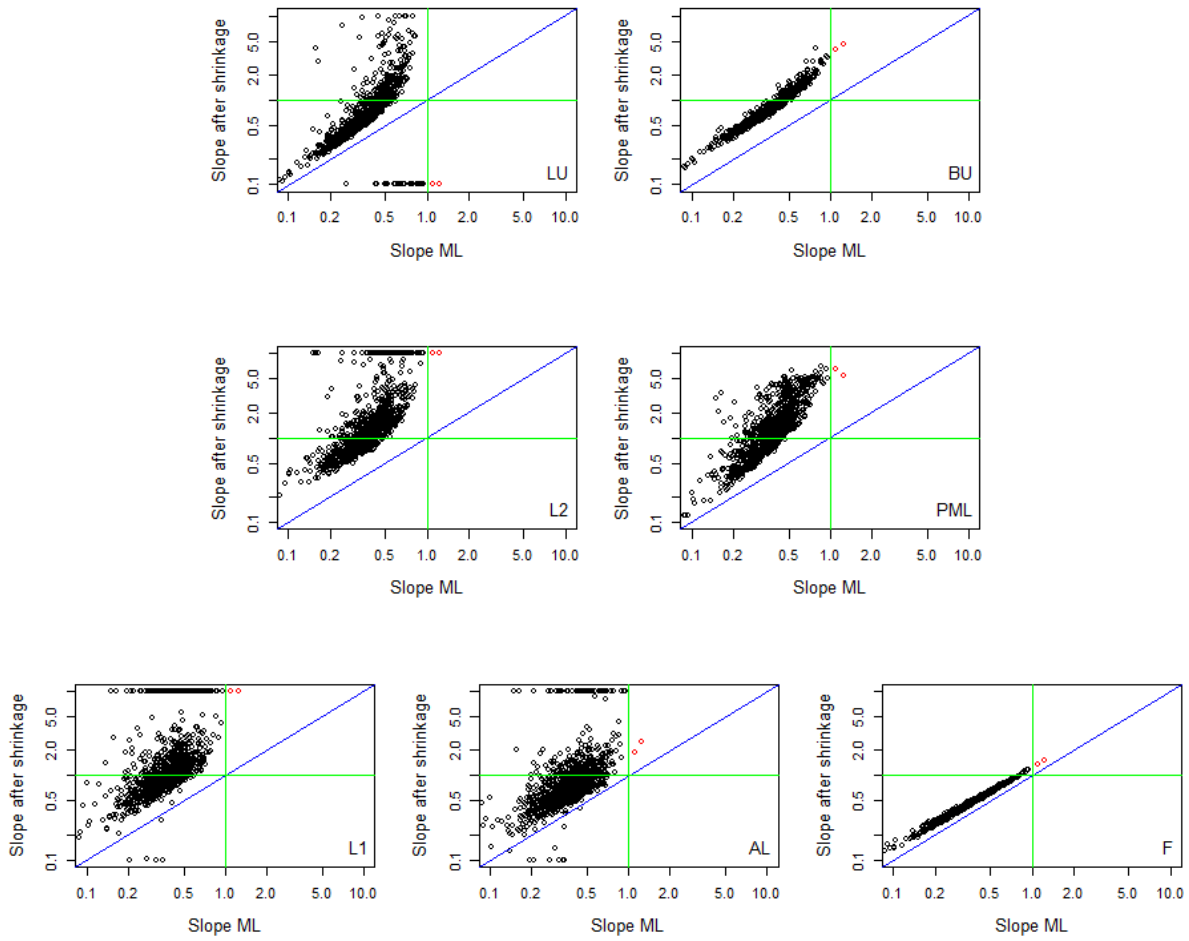
view

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

XXX. 5 true and 5 noise predictors, 0.5 correlation, 10% event rate, 50 EPV

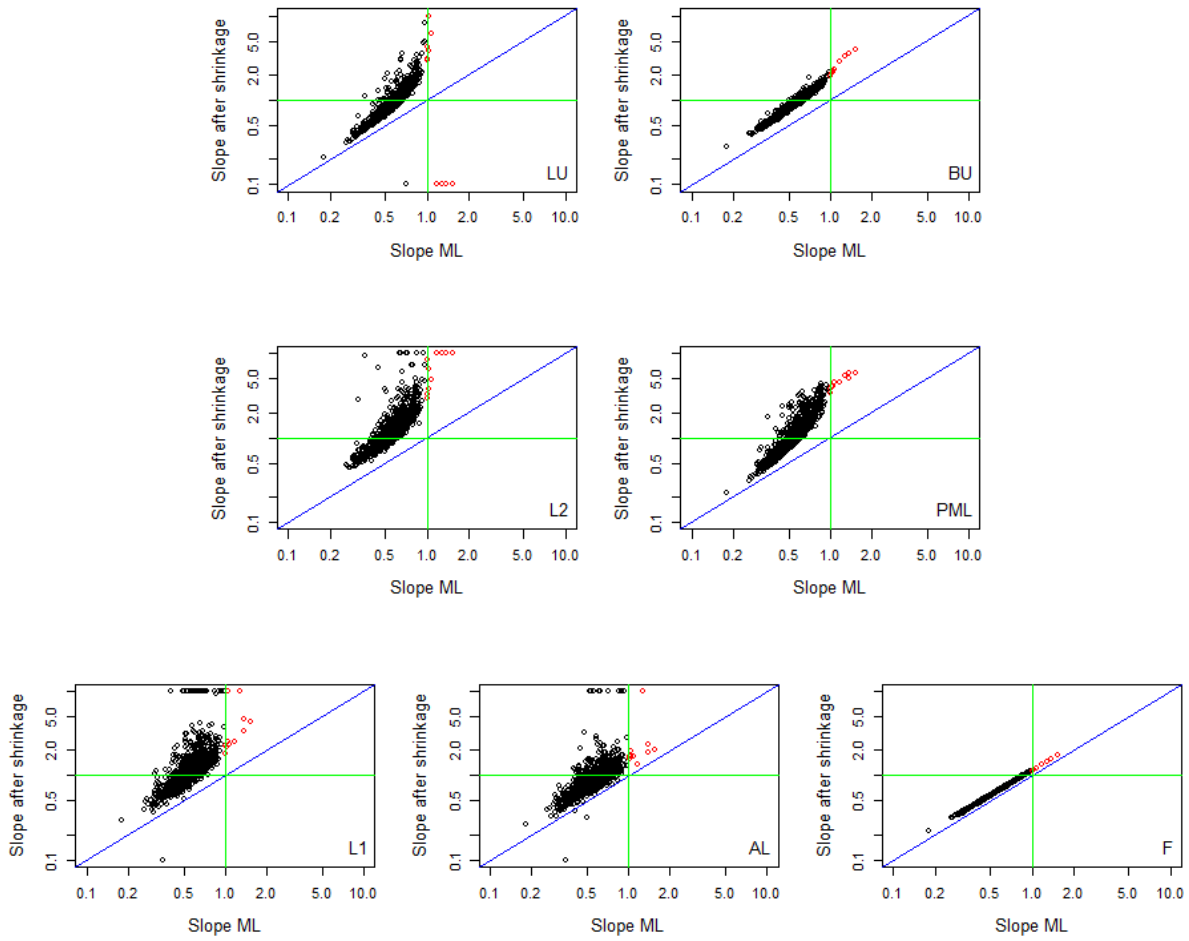


XXXI. 5 true and 5 noise predictors, 0 correlation, 50% event rate, 3 EPV



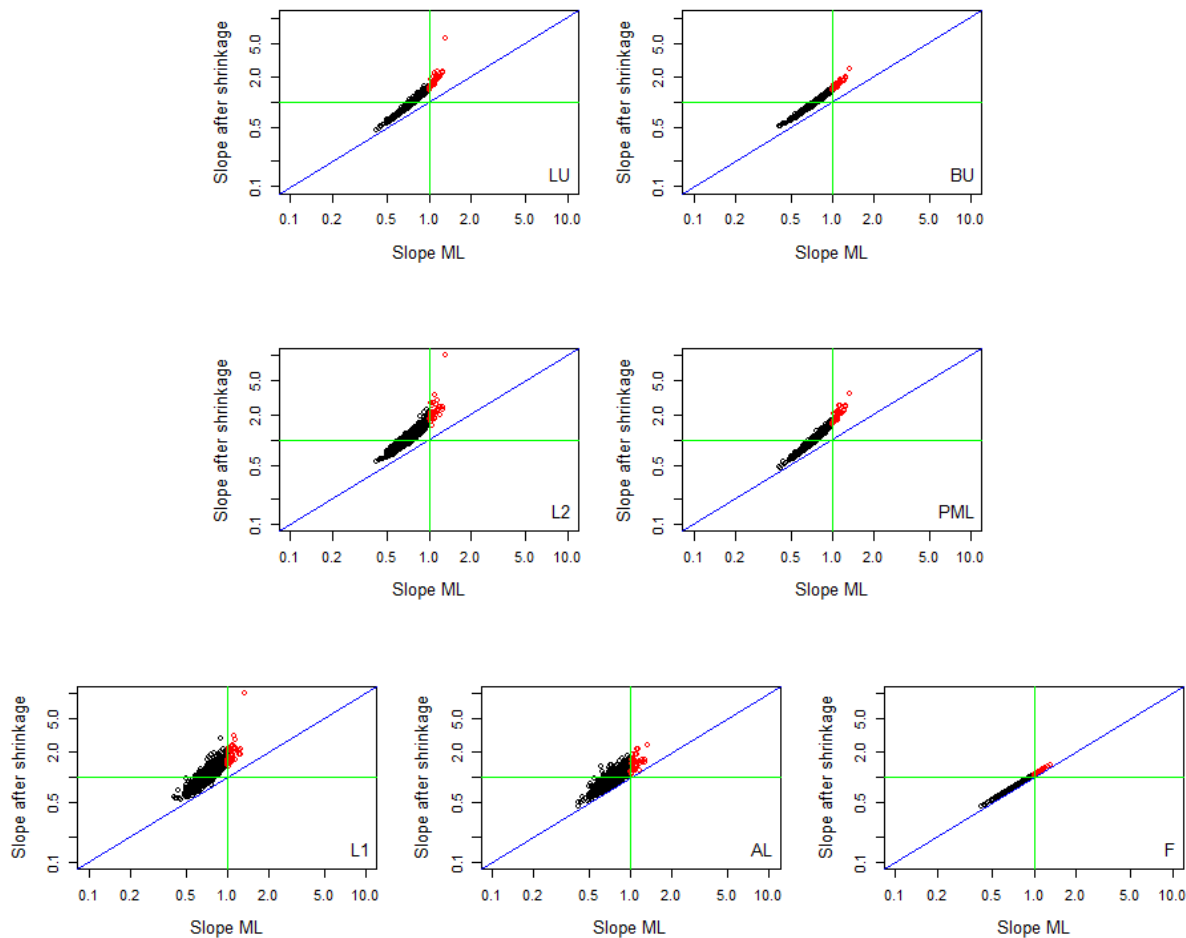
view

XXXII. 5 true and 5 noise predictors, 0 correlation, 50% event rate, 5 EPV



view

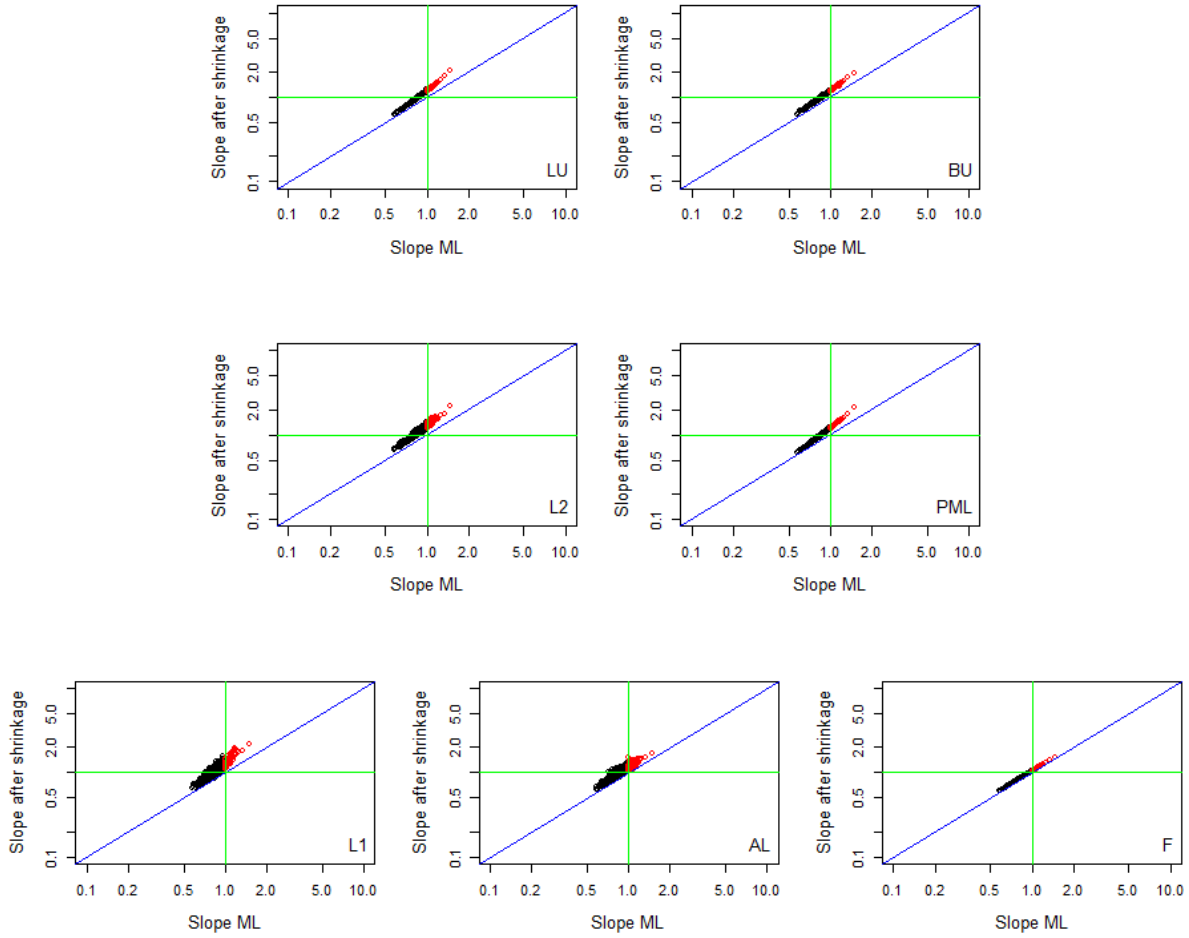
XXXIII. 5 true and 5 noise predictors, 0 correlation, 50% event rate, 10 EPV



view

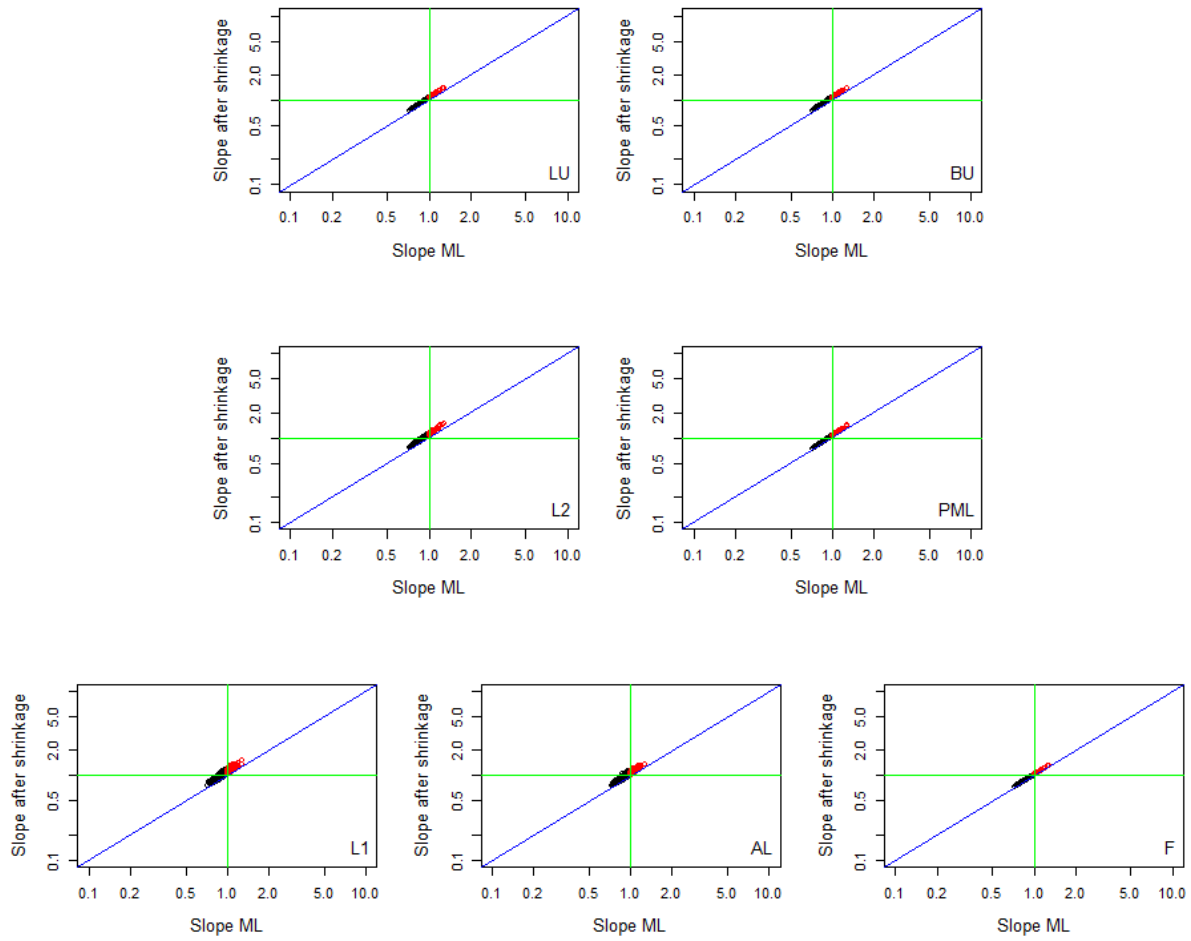
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

XXXIV. 5 true and 5 noise predictors, 0 correlation, 50% event rate, 20 EPV



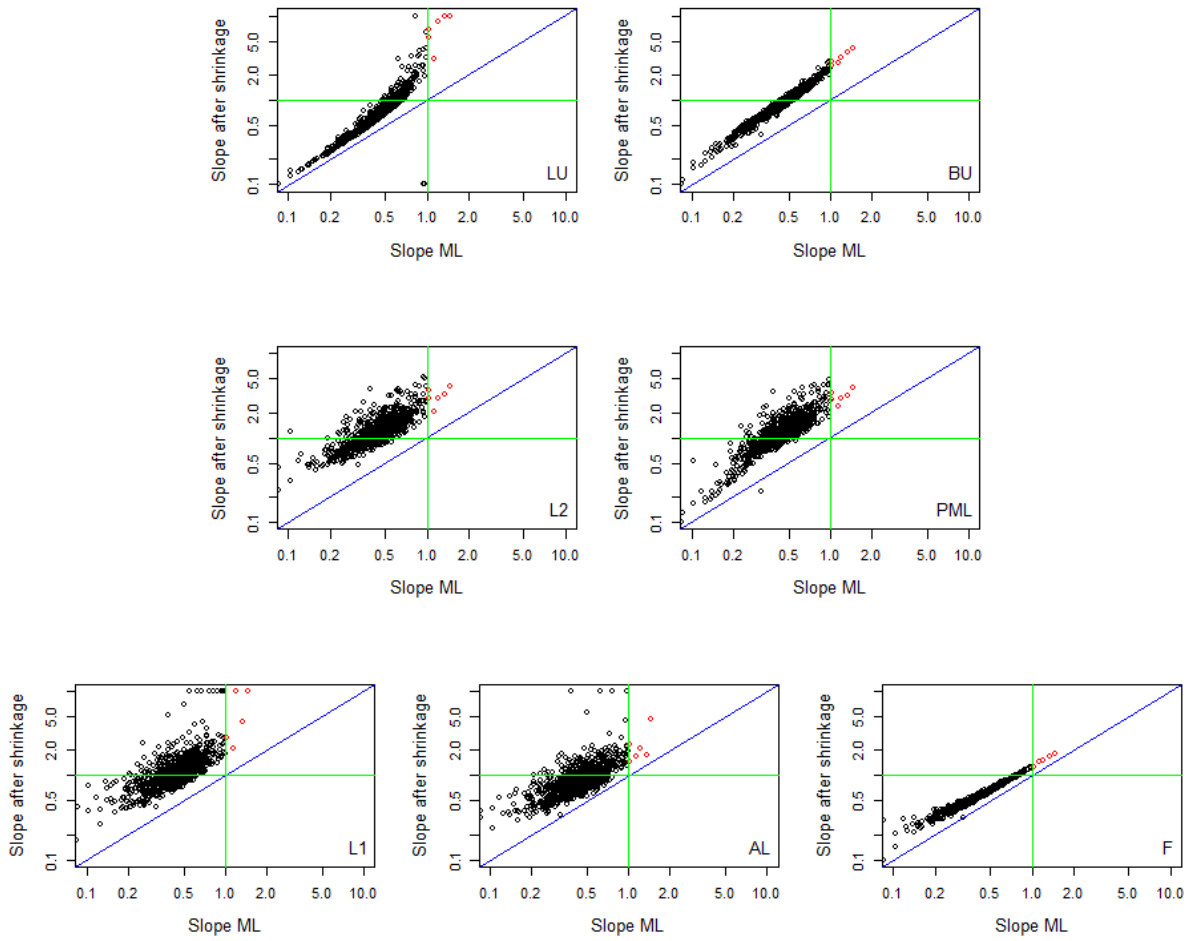
view

XXXV. 5 true and 5 noise predictors, 0 correlation, 50% event rate, 50 EPV



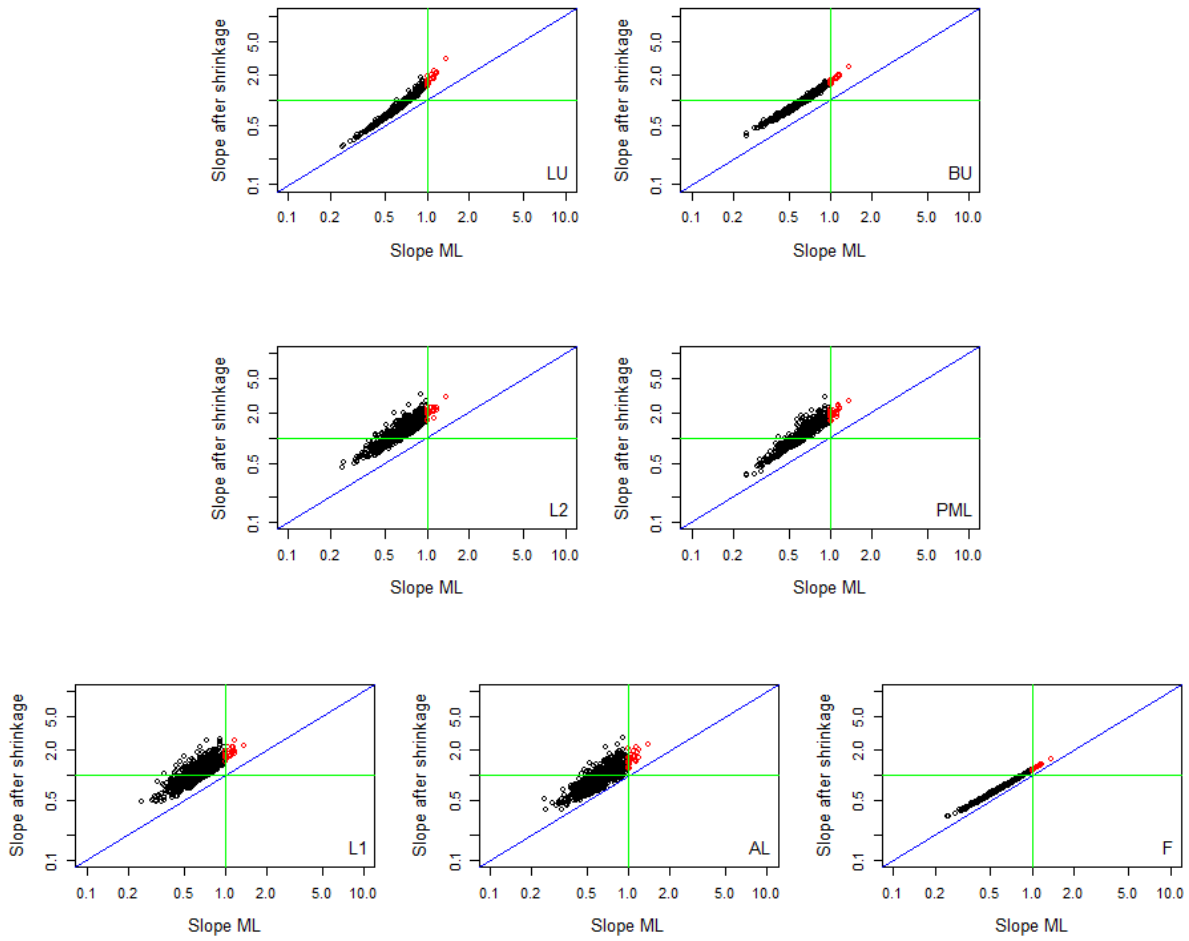
view

XXXVI. 5 true and 5 noise predictors, 0.5 correlation, 50% event rate, 3 EPV



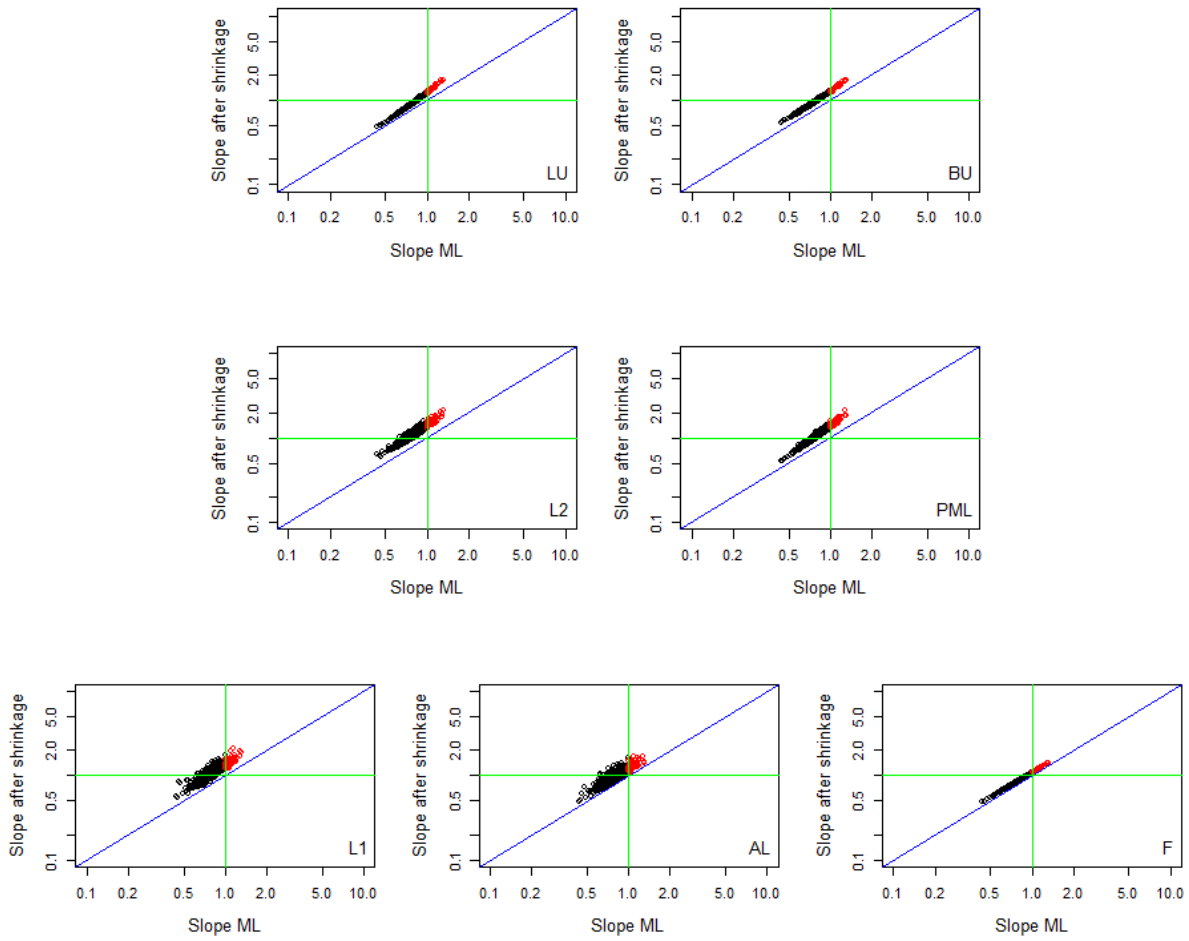
view

XXXVII. 5 true and 5 noise predictors, 0.5 correlation, 50% event rate, 5 EPV



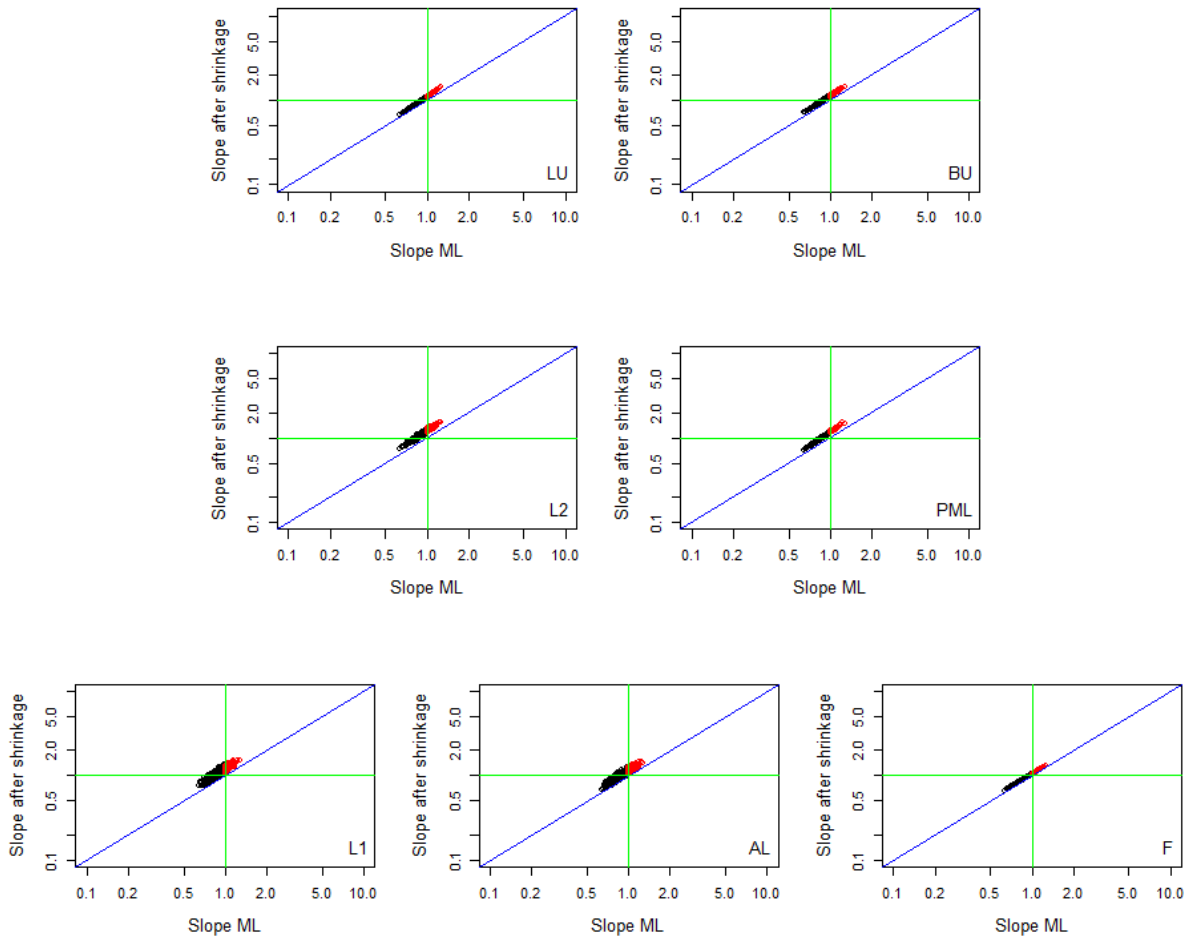
view

XXXVIII. 5 true and 5 noise predictors, 0.5 correlation, 50% event rate, 10 EPV



view

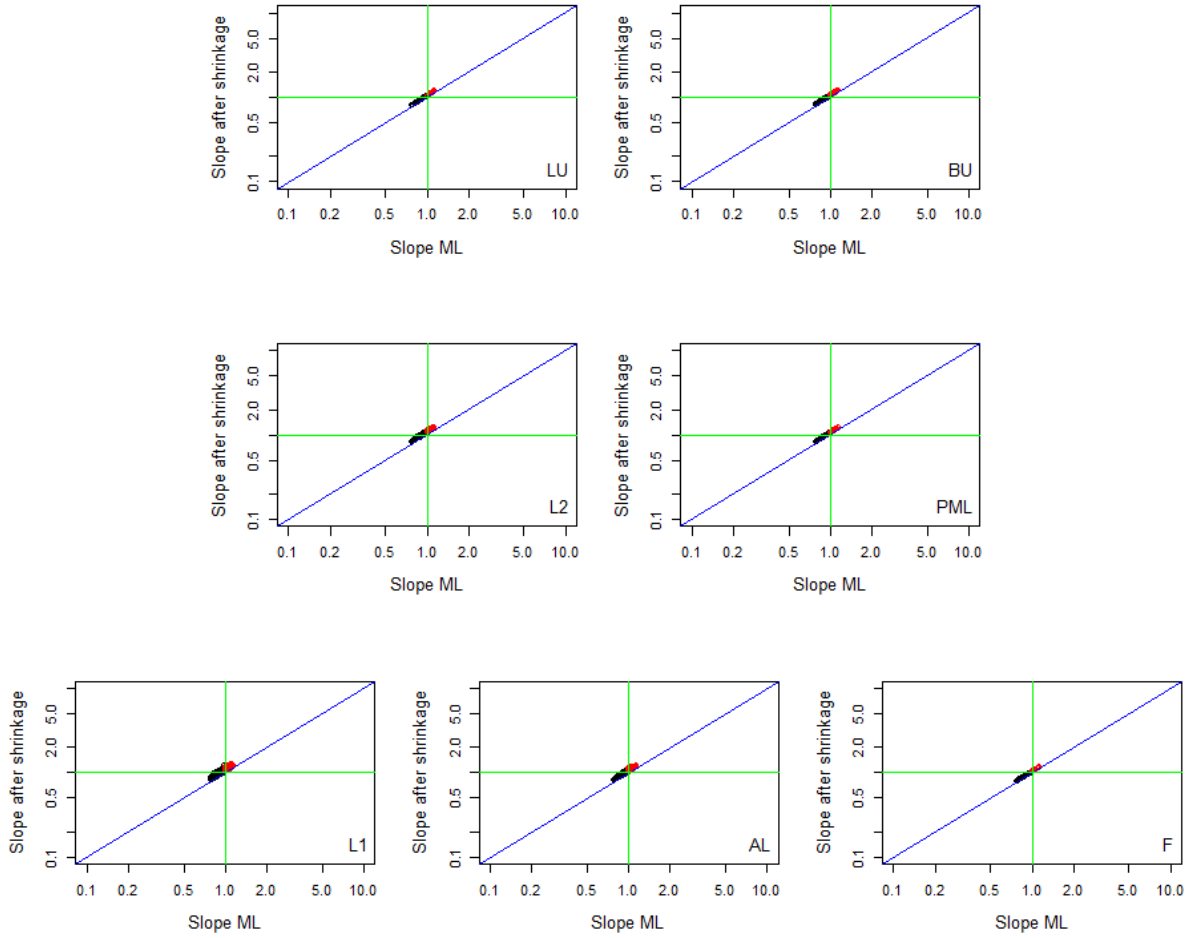
XXXIX. 5 true and 5 noise predictors, 0.5 correlation, 50% event rate, 20 EPV



view

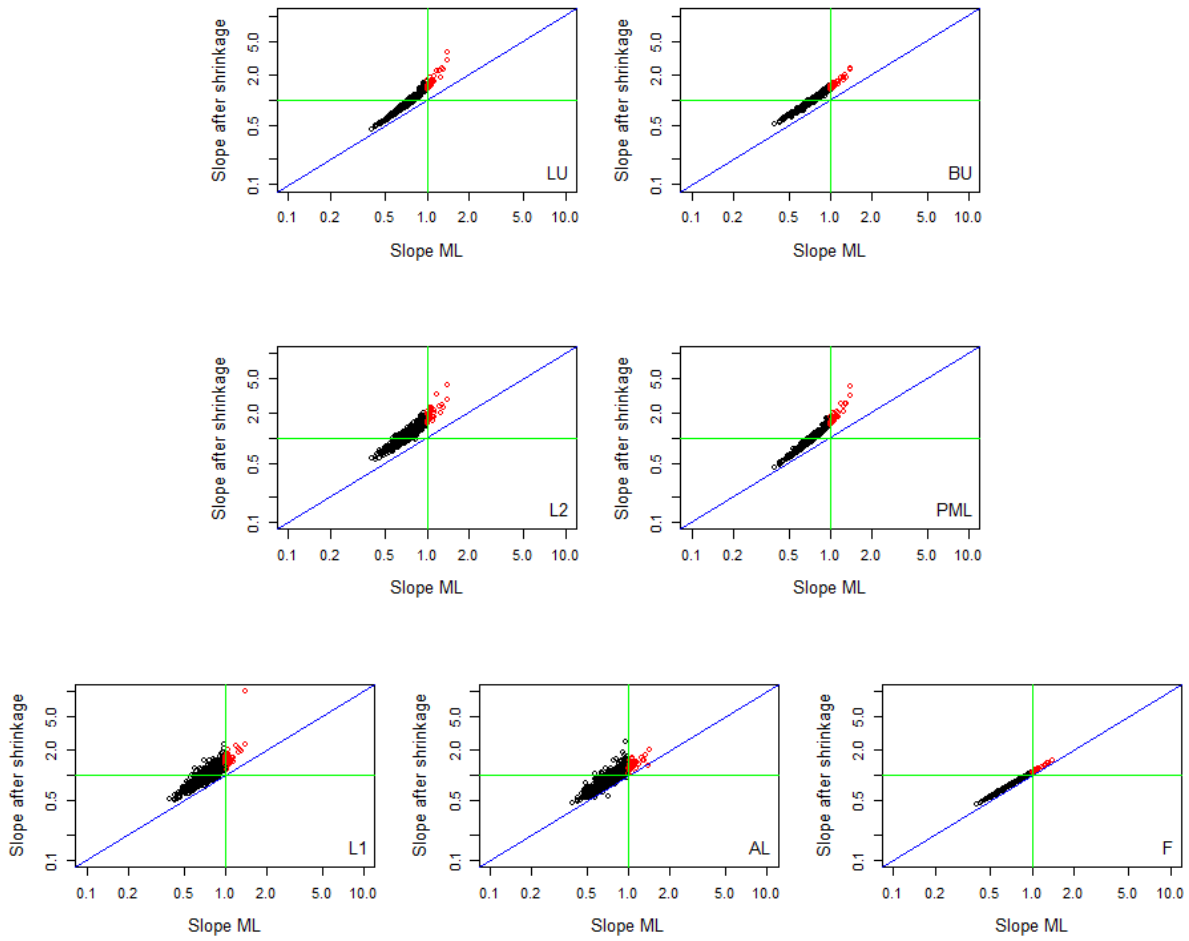
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

XL. 5 true and 5 noise predictors, 0.5 correlation, 50% event rate, 50 EPV



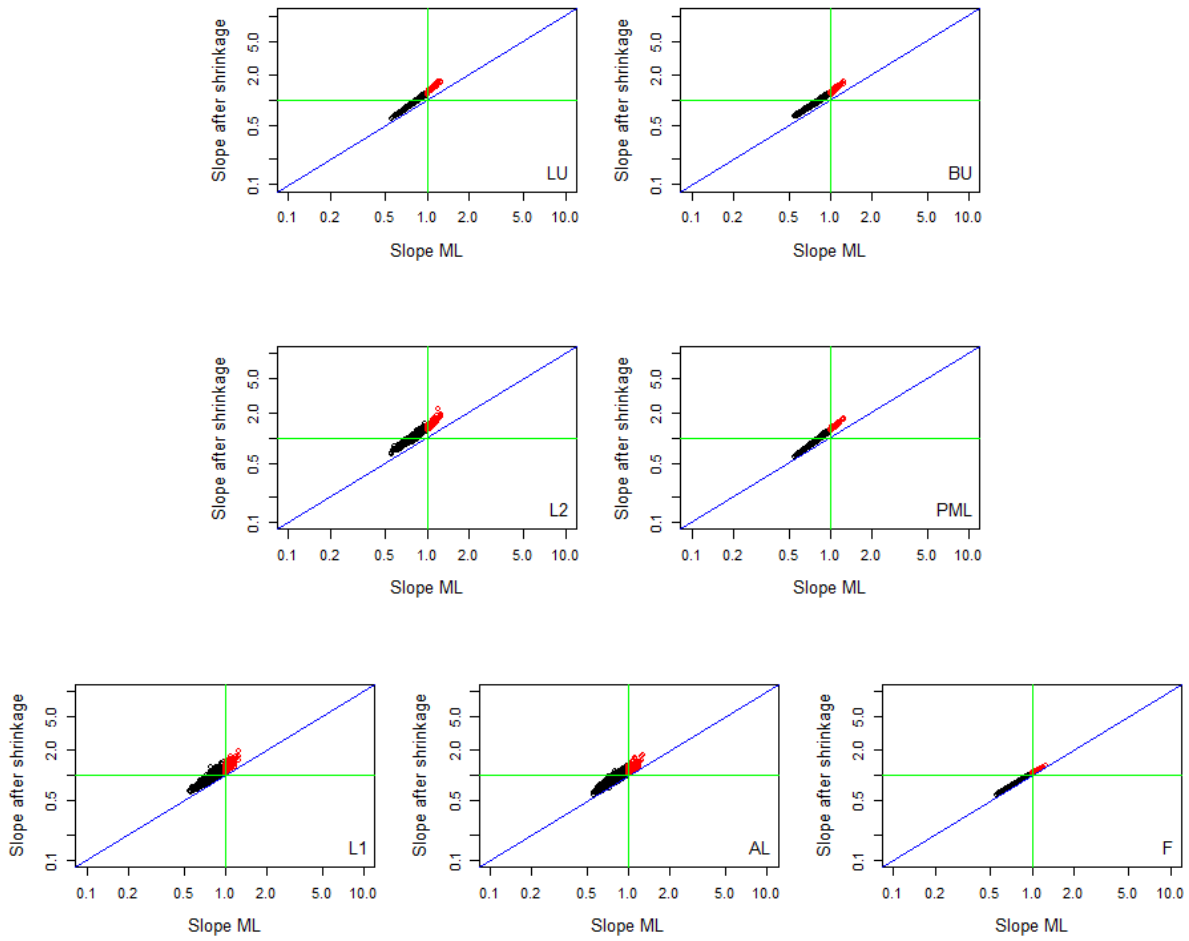
view

XLI. 10 true predictors, 0 correlation, 10% event rate, 3 EPV



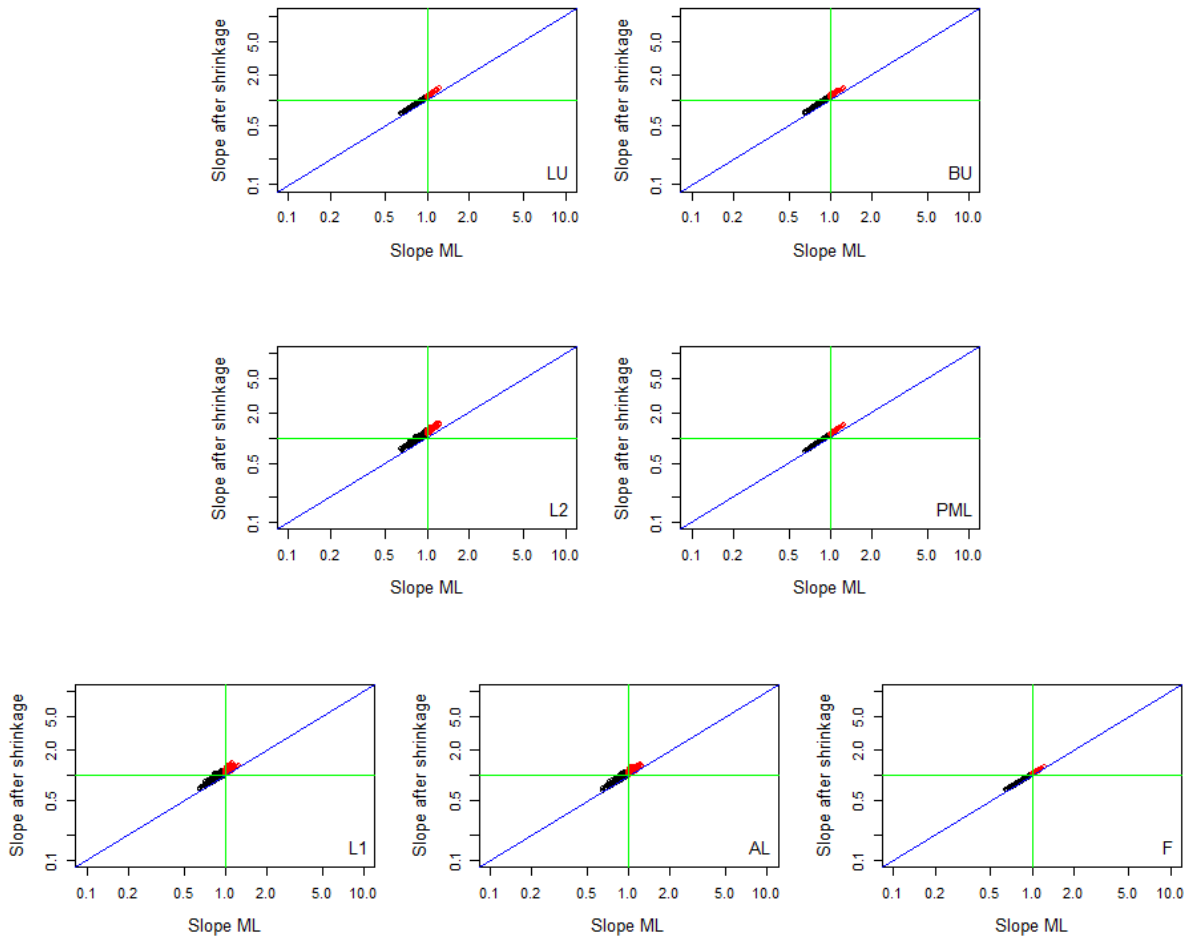
view

XLII. 10 true predictors, 0 correlation, 10% event rate, 5 EPV



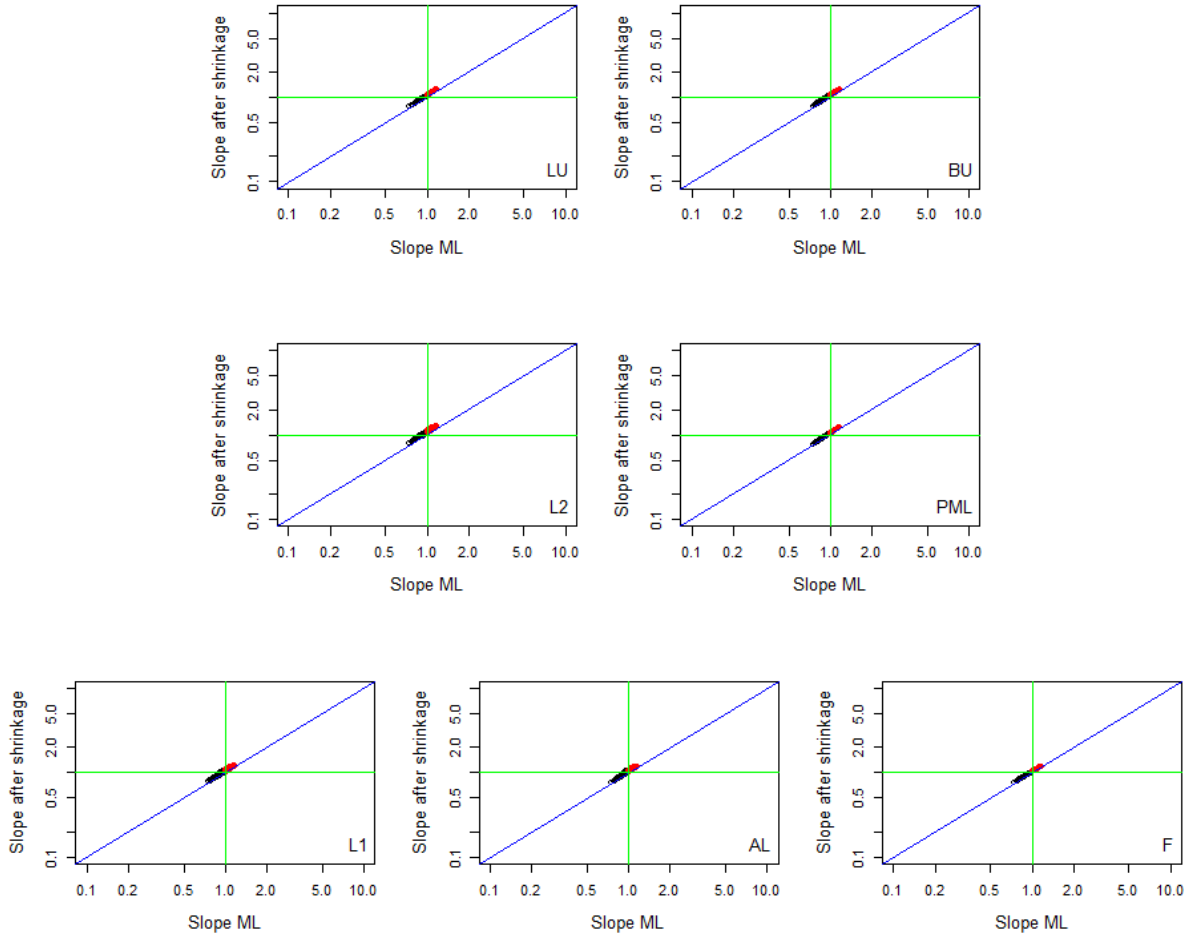
view

XLIII. 10 true predictors, 0 correlation, 10% event rate, 10 EPV



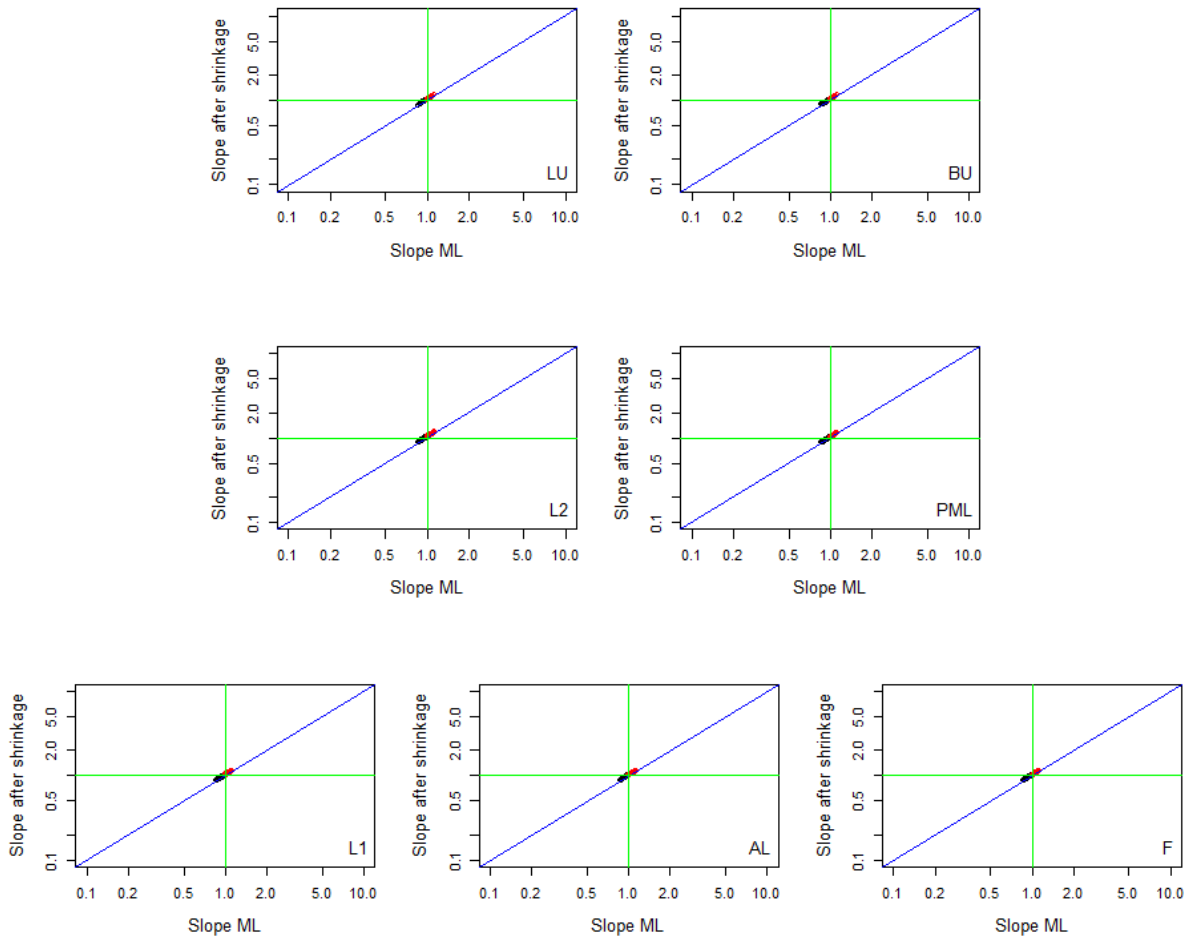
view

XLIV. 10 true predictors, 0 correlation, 10% event rate, 20 EPV

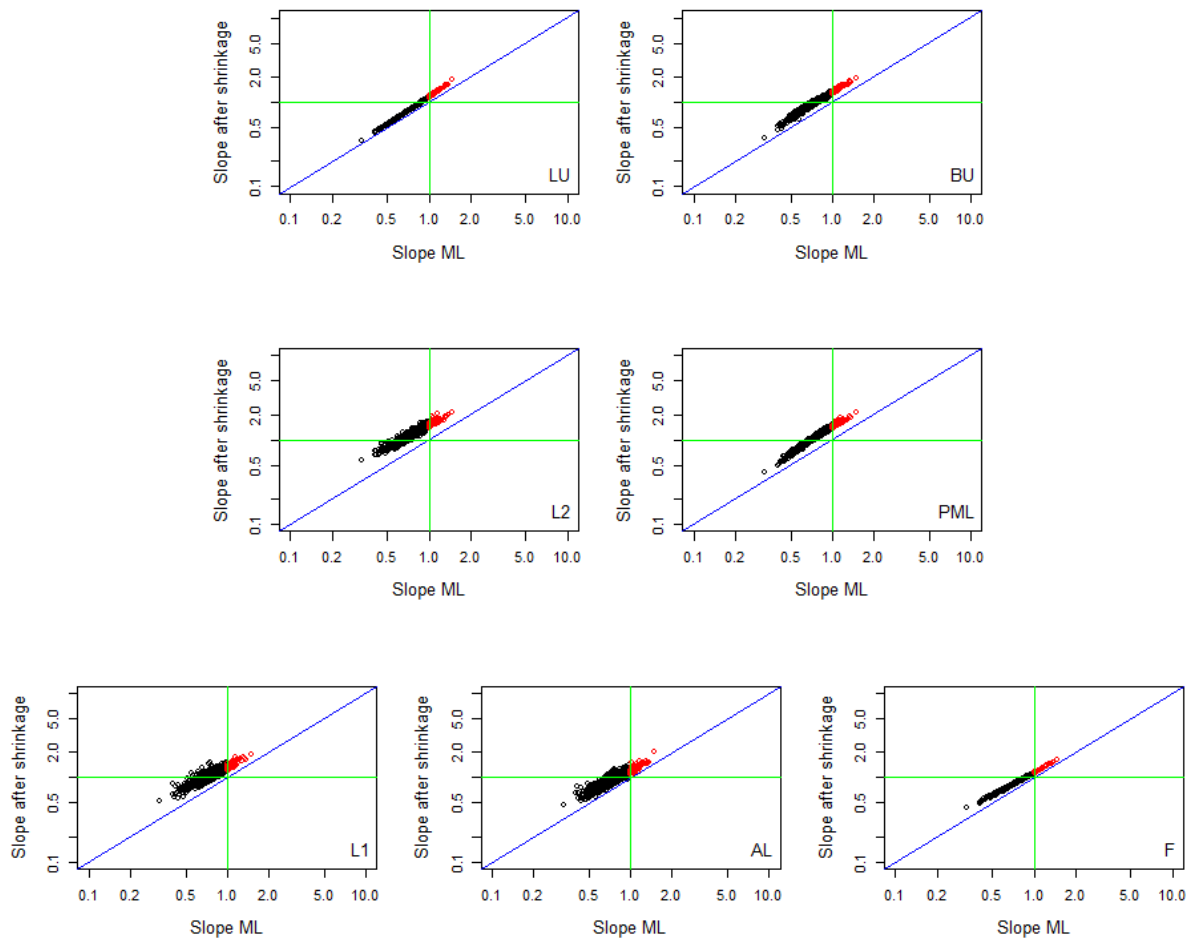


view

XLV. 10 true predictors, 0 correlation, 10% event rate, 50 EPV

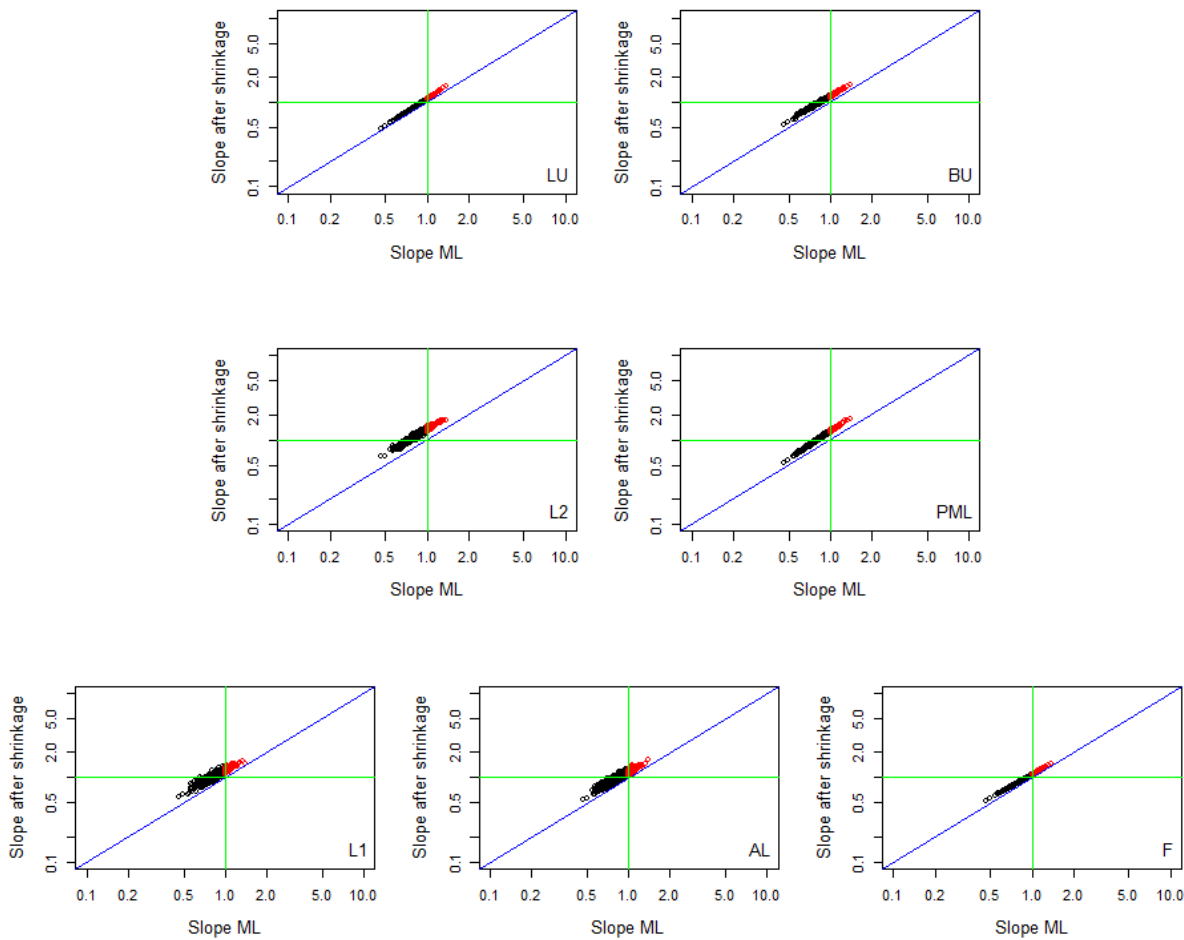


XLVI. 10 true predictors, 0.5 correlation, 10% event rate, 3 EPV



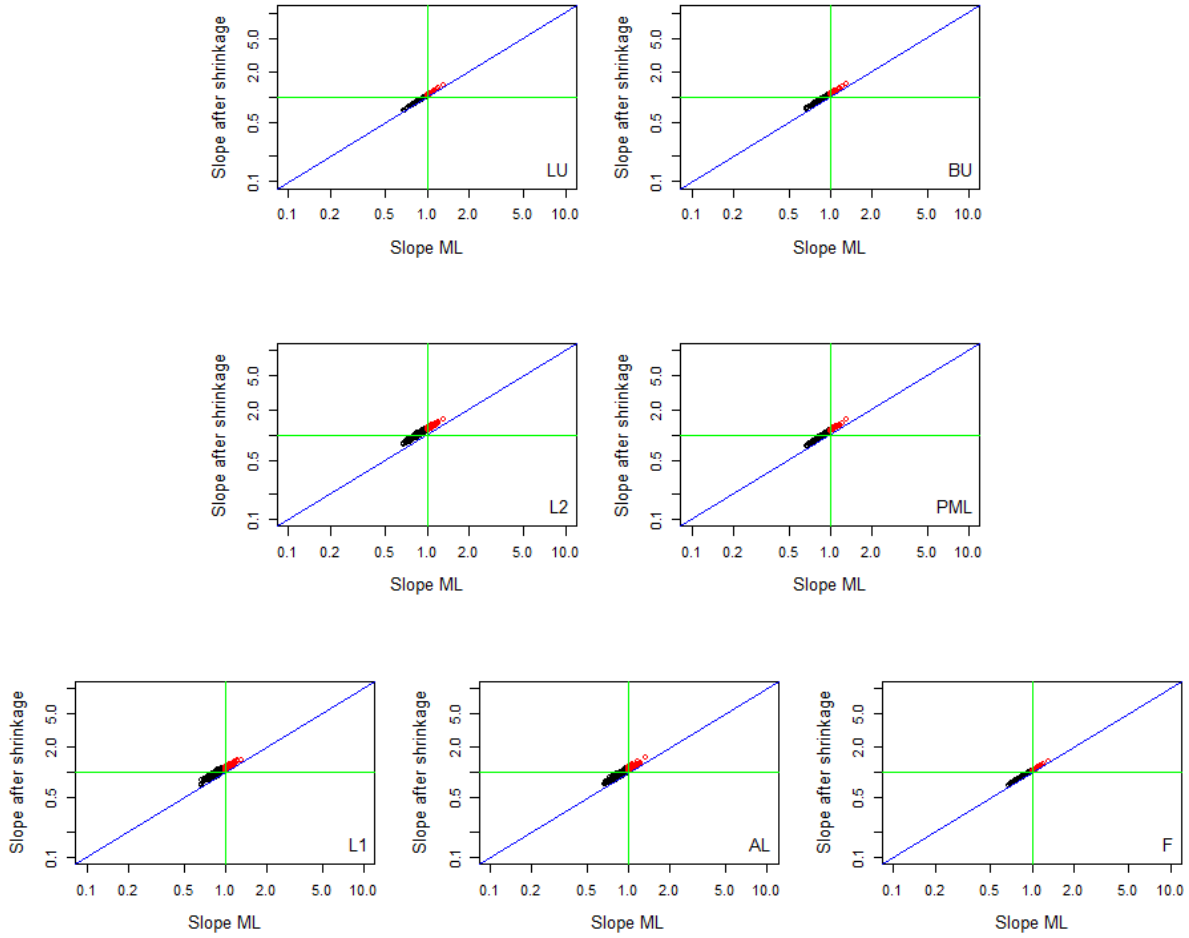
ew

XLVII. 10 true predictors, 0.5 correlation, 10% event rate, 5 EPV



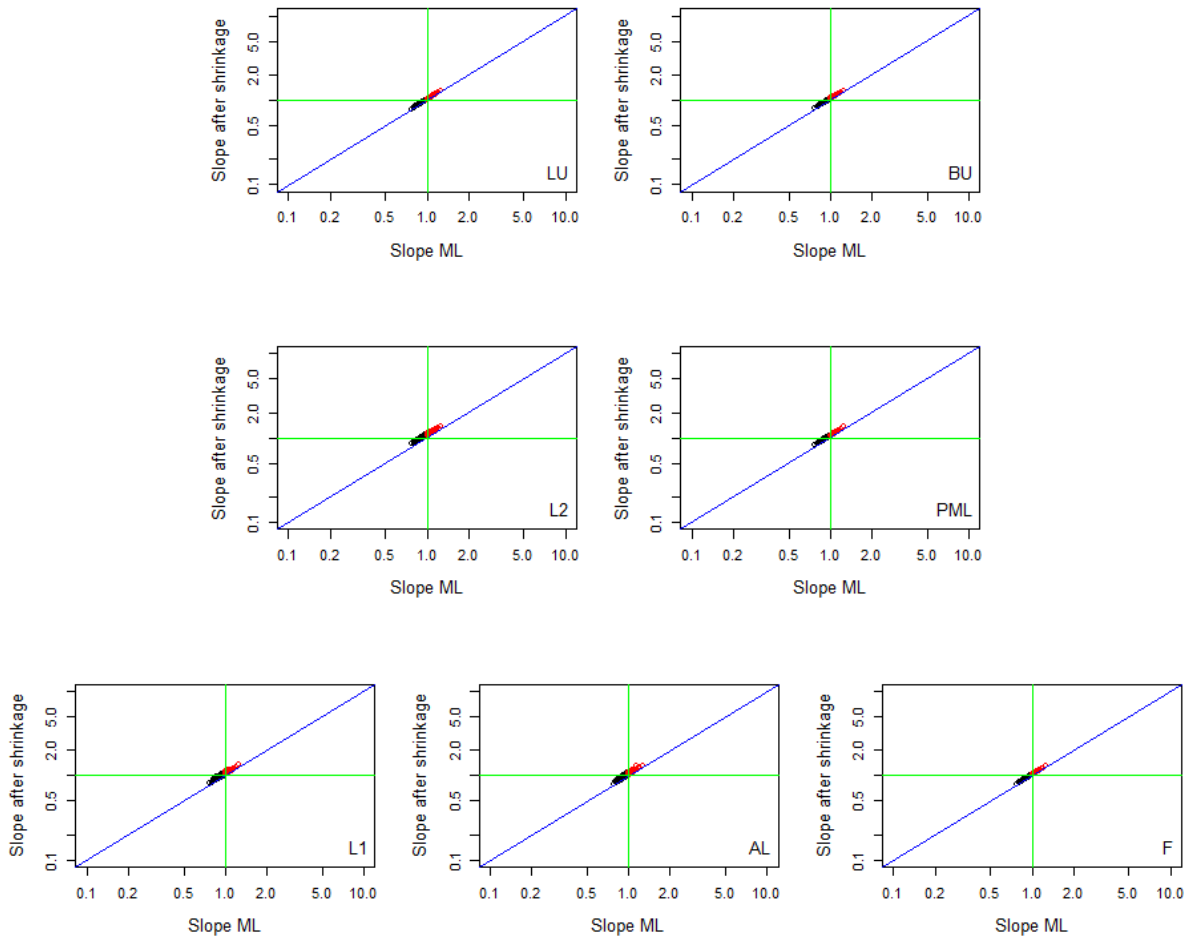
view

XLVIII. 10 true predictors, 0.5 correlation, 10% event rate, 10 EPV



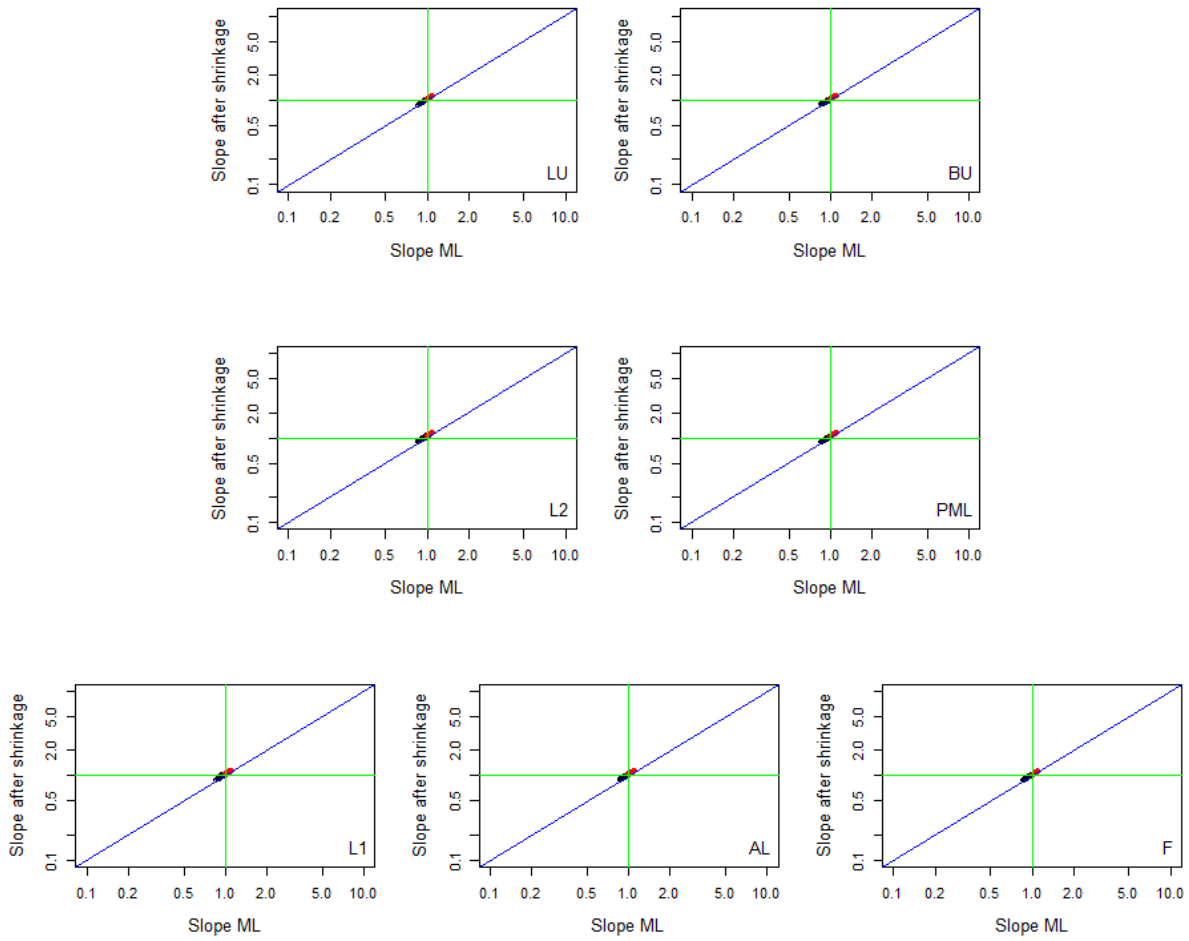
view

XLIX. 10 true predictors, 0.5 correlation, 10% event rate, 20 EPV



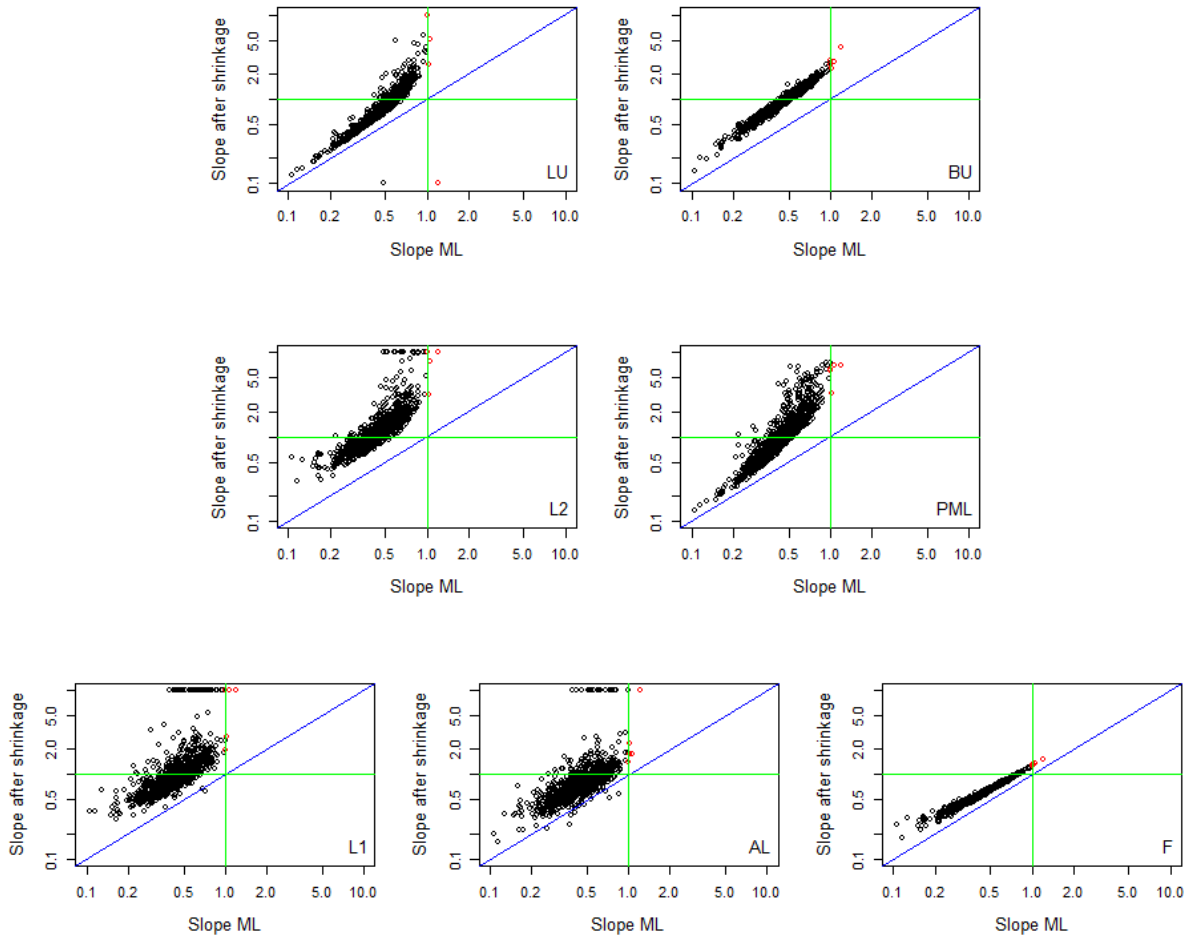
view

L. 10 true predictors, 0.5 correlation, 10% event rate, 50 EPV



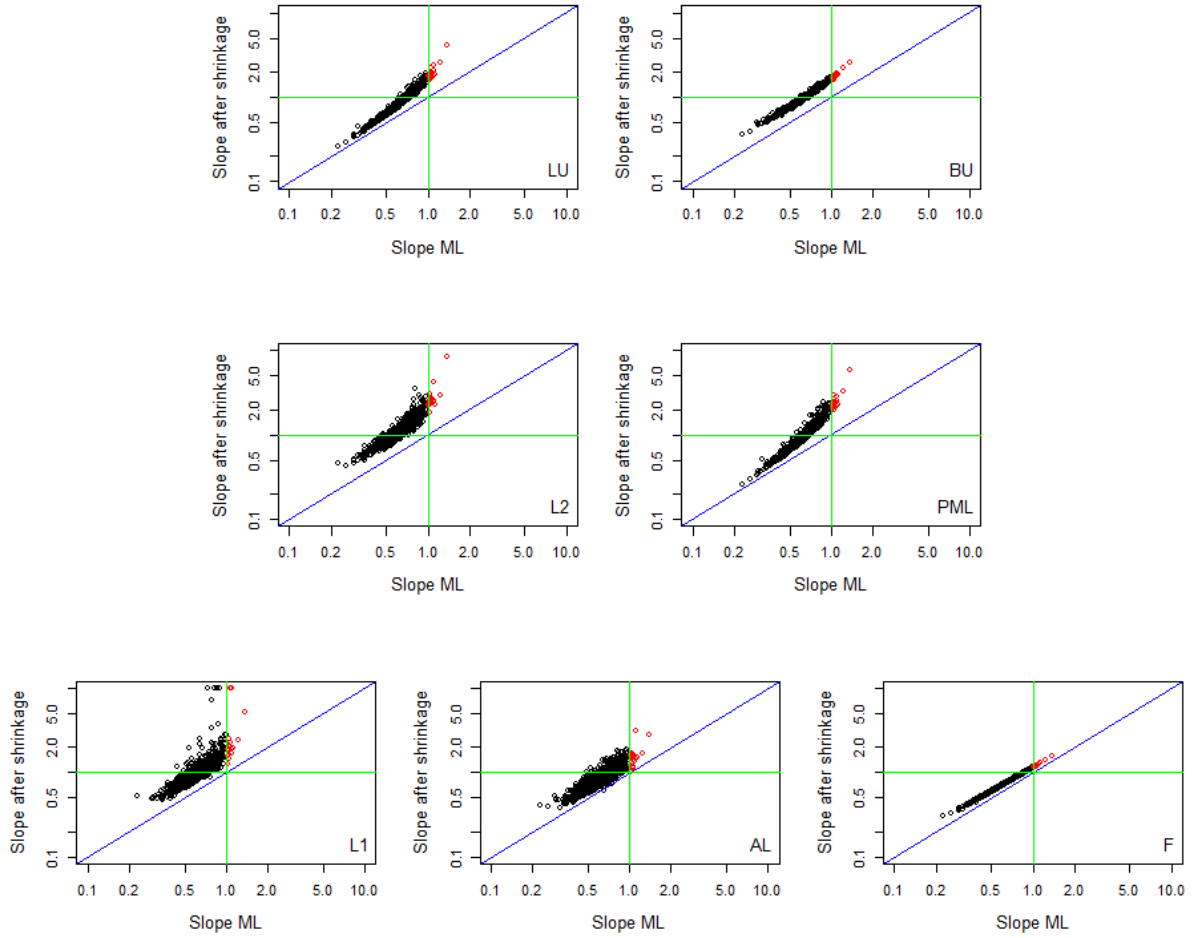
view

LI. 10 true predictors, 0 correlation, 50% event rate, 3 EPV



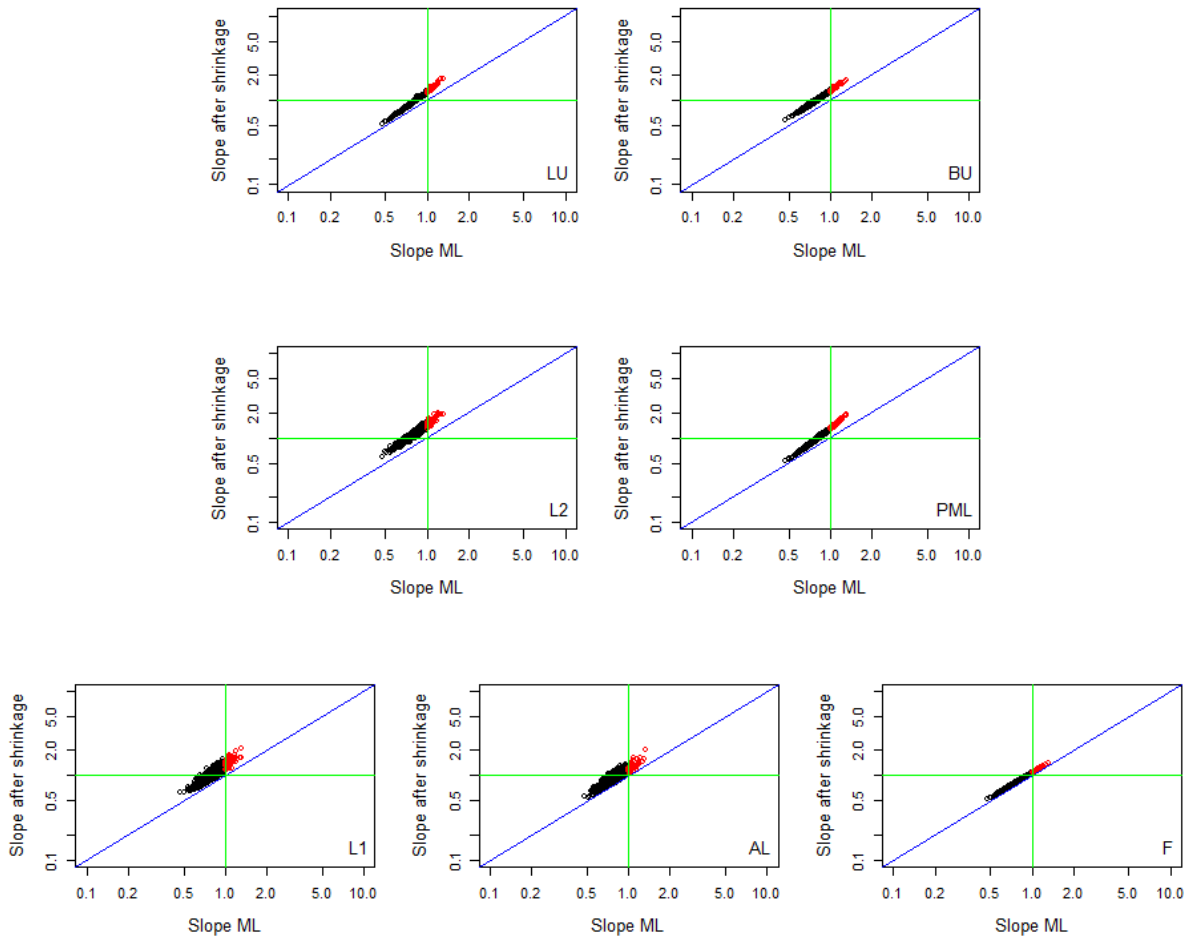
ew

LII. 10 true predictors, 0 correlation, 50% event rate, 5 EPV



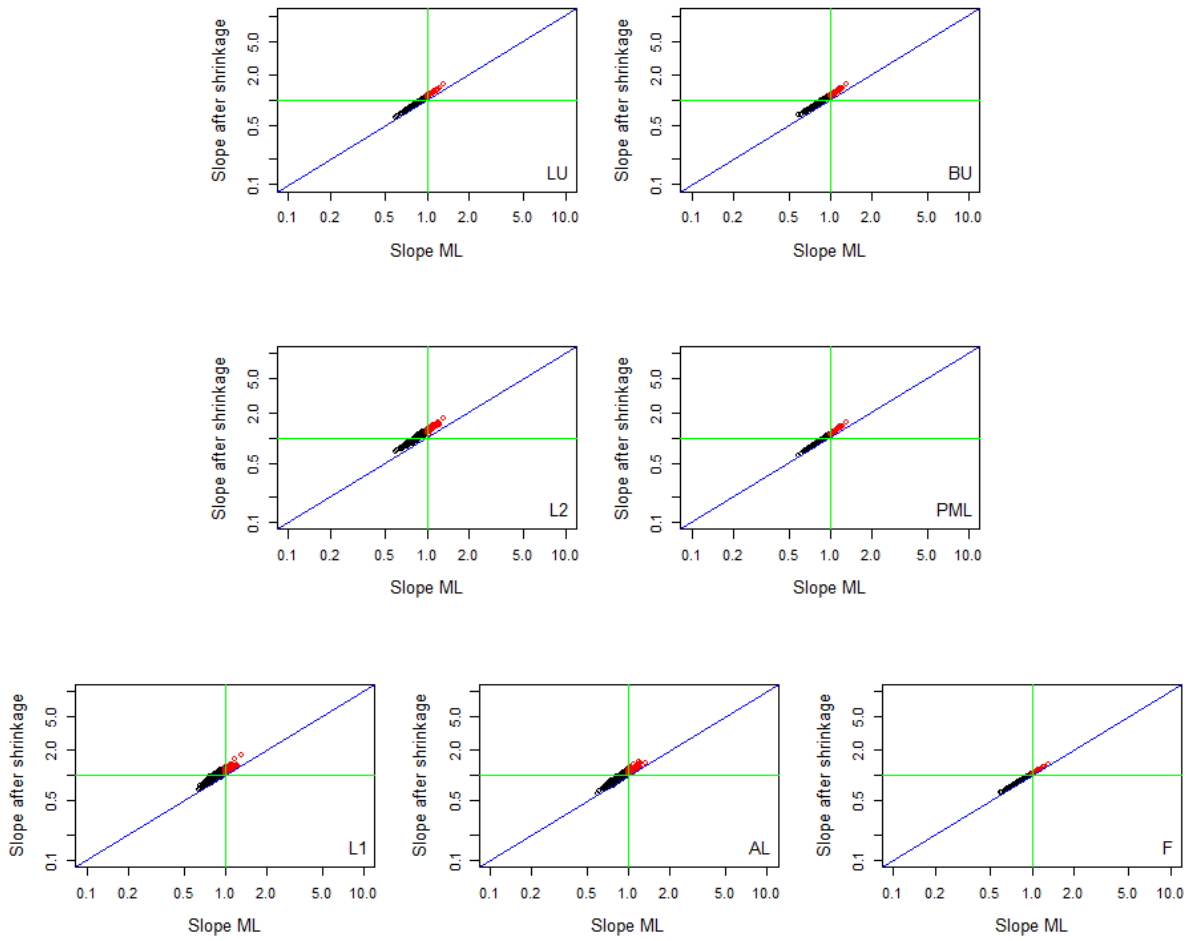
view

LIII. 10 true predictors, 0 correlation, 50% event rate, 10 EPV



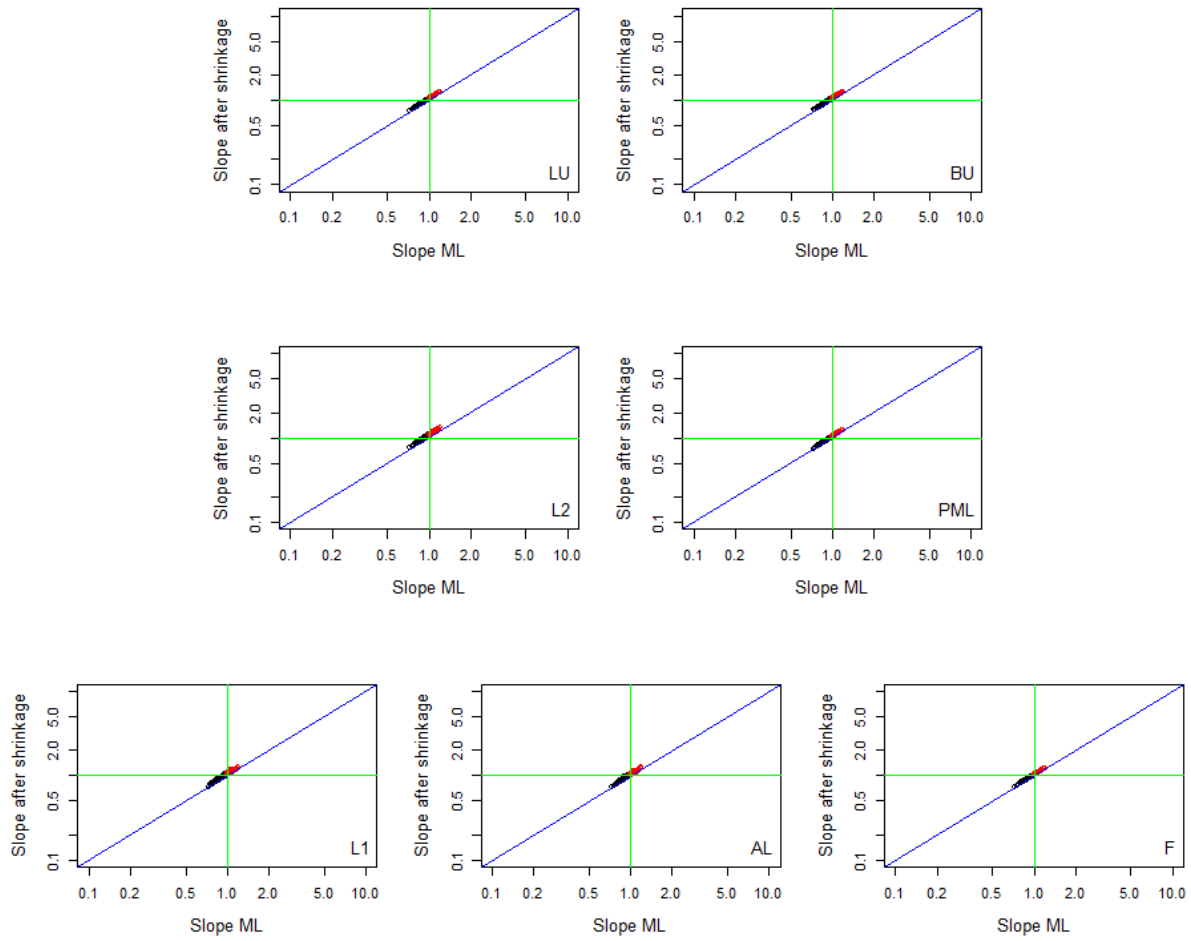
view

LIV. 10 true predictors, 0 correlation, 50% event rate, 20 EPV



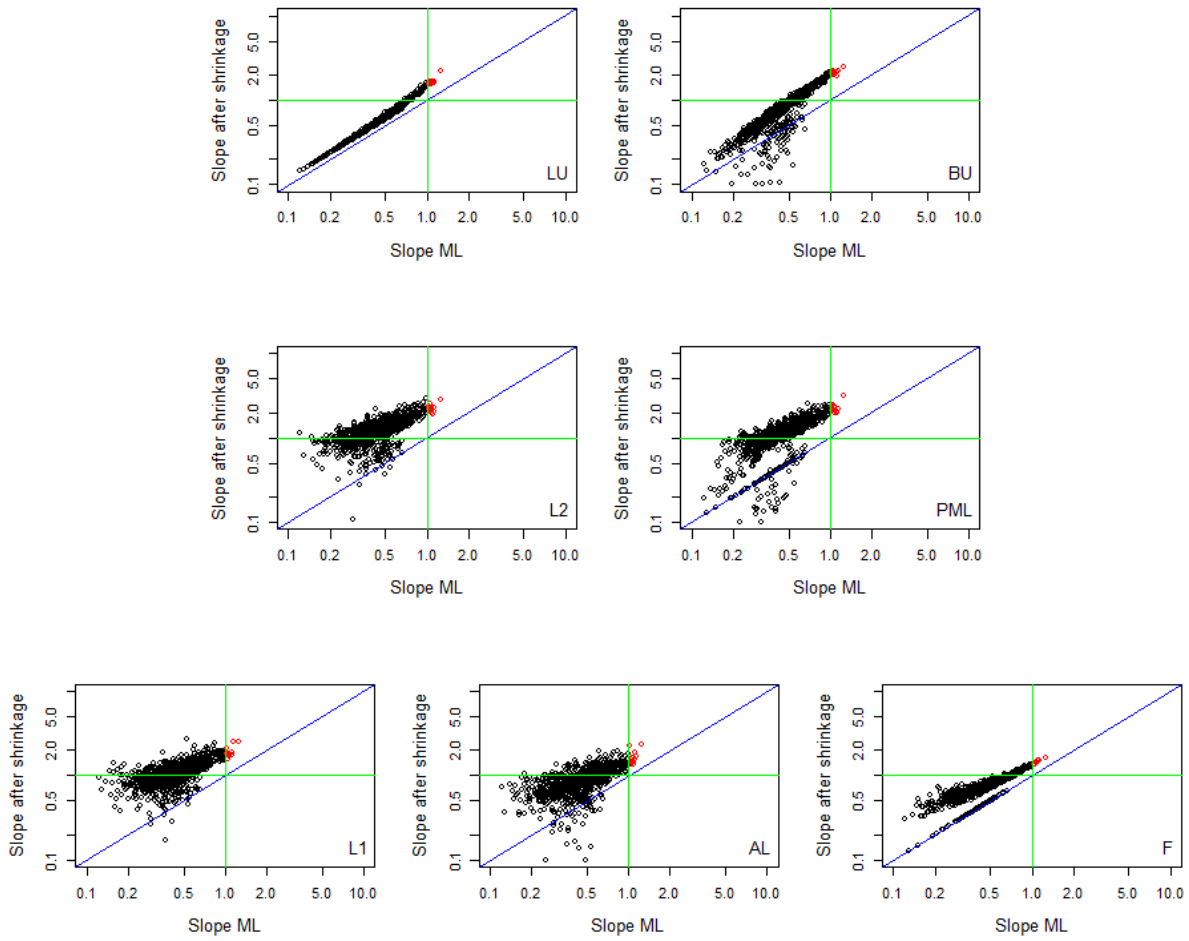
view

LV. 10 true predictors, 0 correlation, 50% event rate, 50 EPV



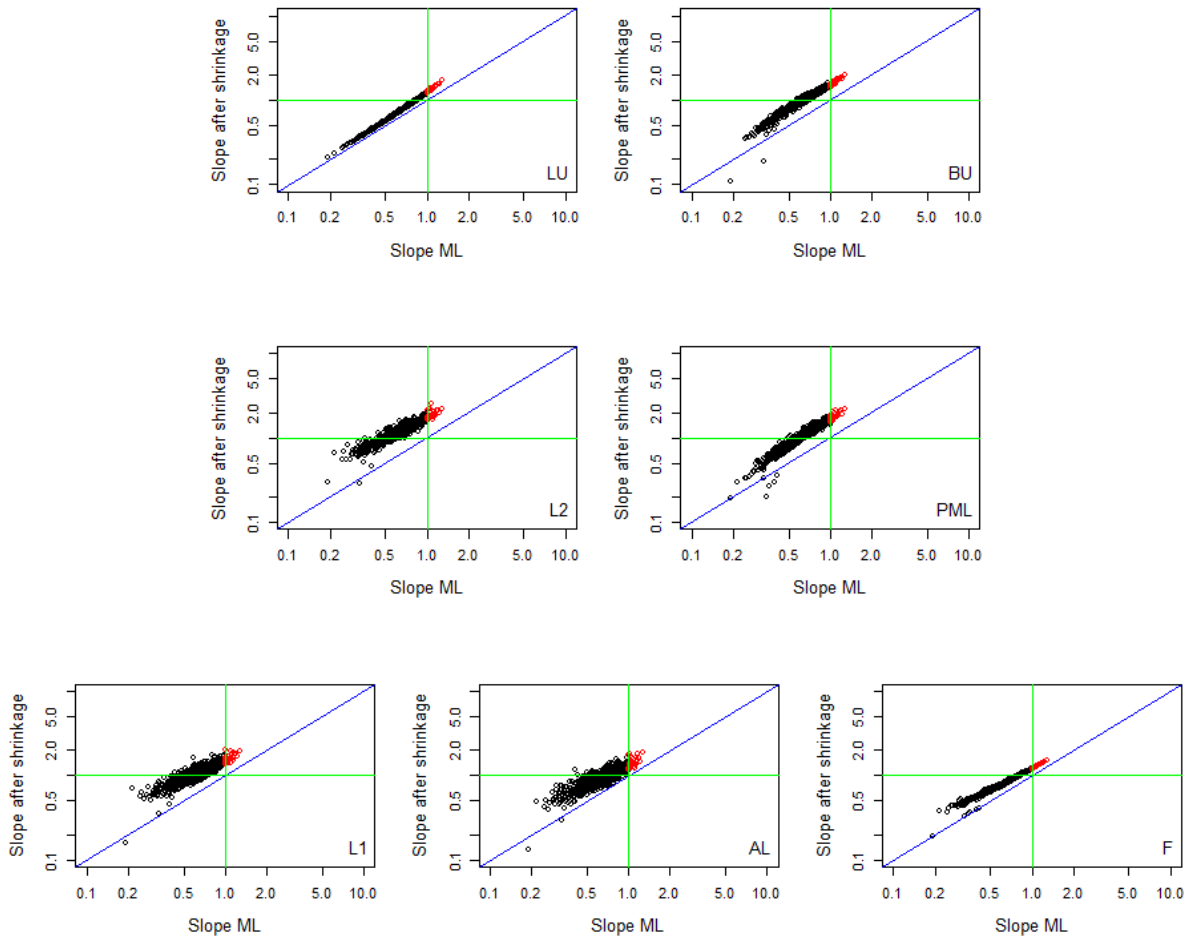
view

LVI. 10 true predictors, 0.5 correlation, 50% event rate, 3 EPV



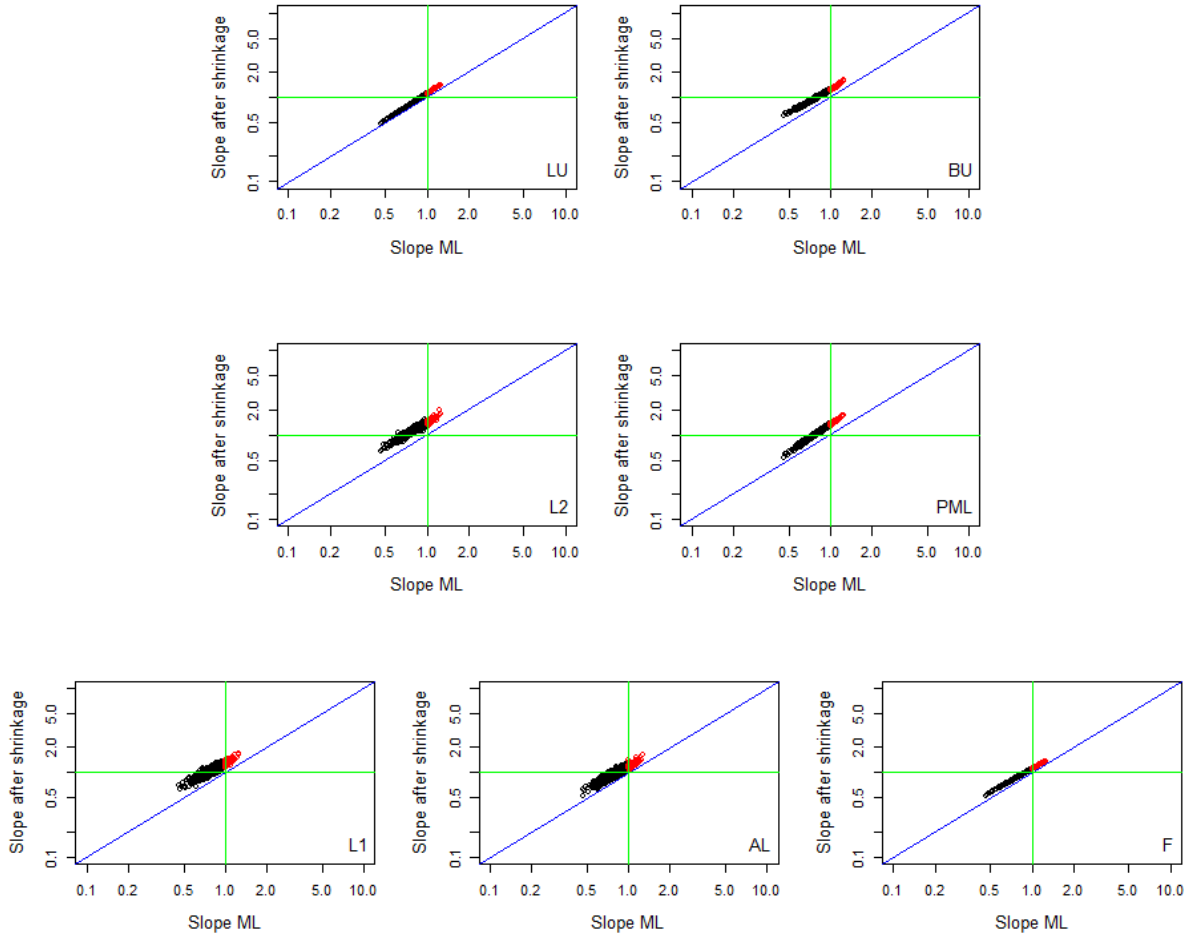
view

LVII. 10 true predictors, 0.5 correlation, 50% event rate, 5 EPV



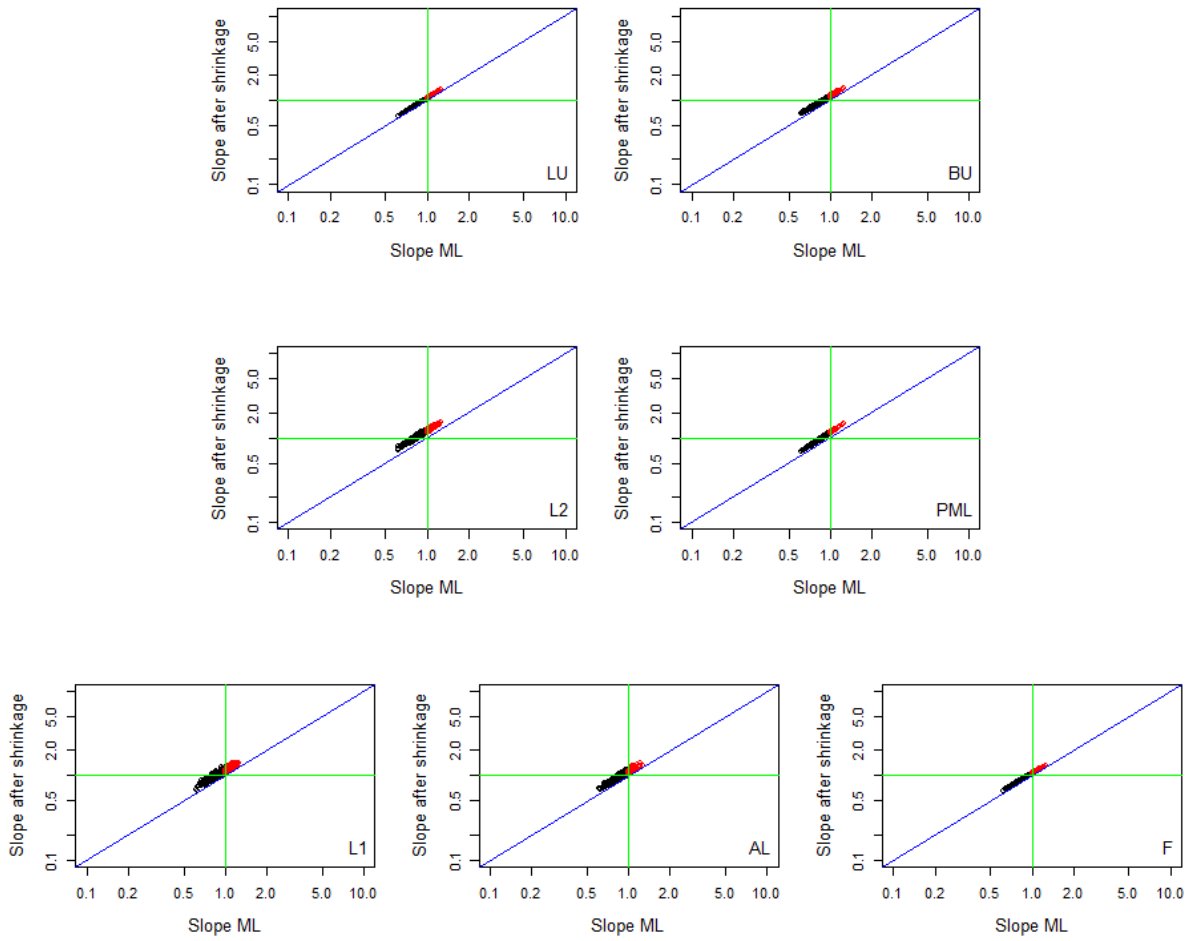
view

LVIII. 10 true predictors, 0.5 correlation, 50% event rate, 10 EPV



view

LIX. 10 true predictors, 0.5 correlation, 50% event rate, 20 EPV



view

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

LX. 10 true predictors, 0.5 correlation, 50% event rate, 50 EPV

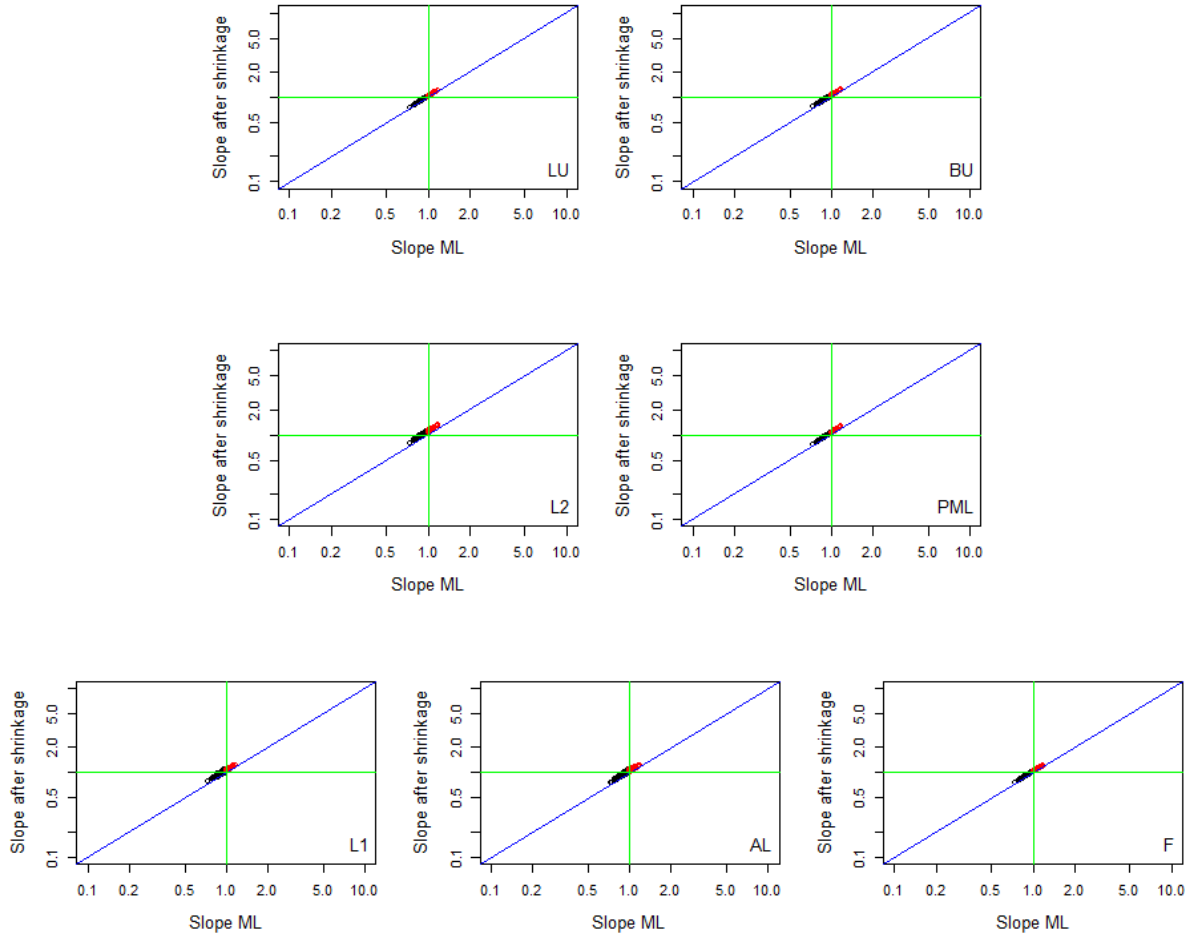
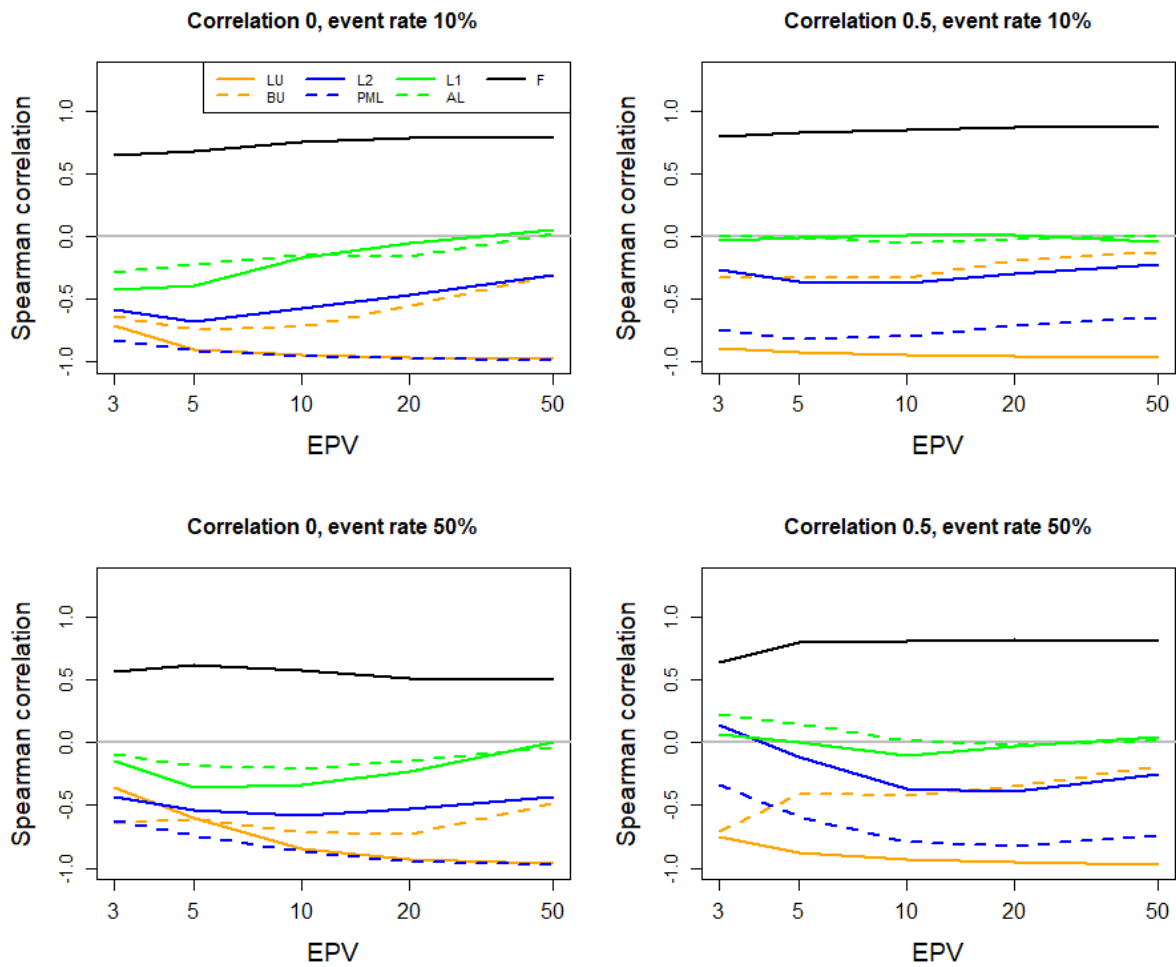
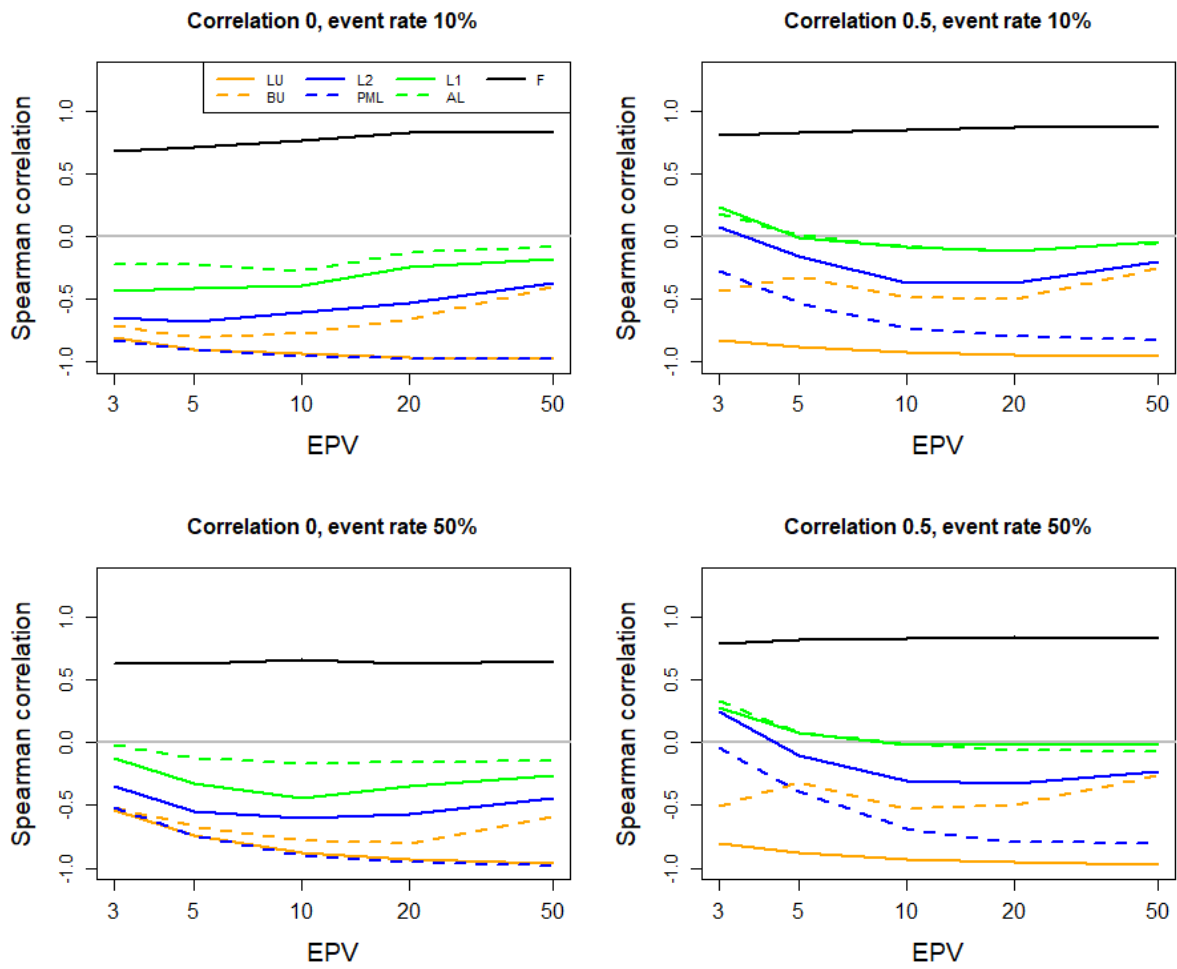


Figure S8. Spearman correlation between estimated and optimal shrinkage by scenario and method. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression ; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

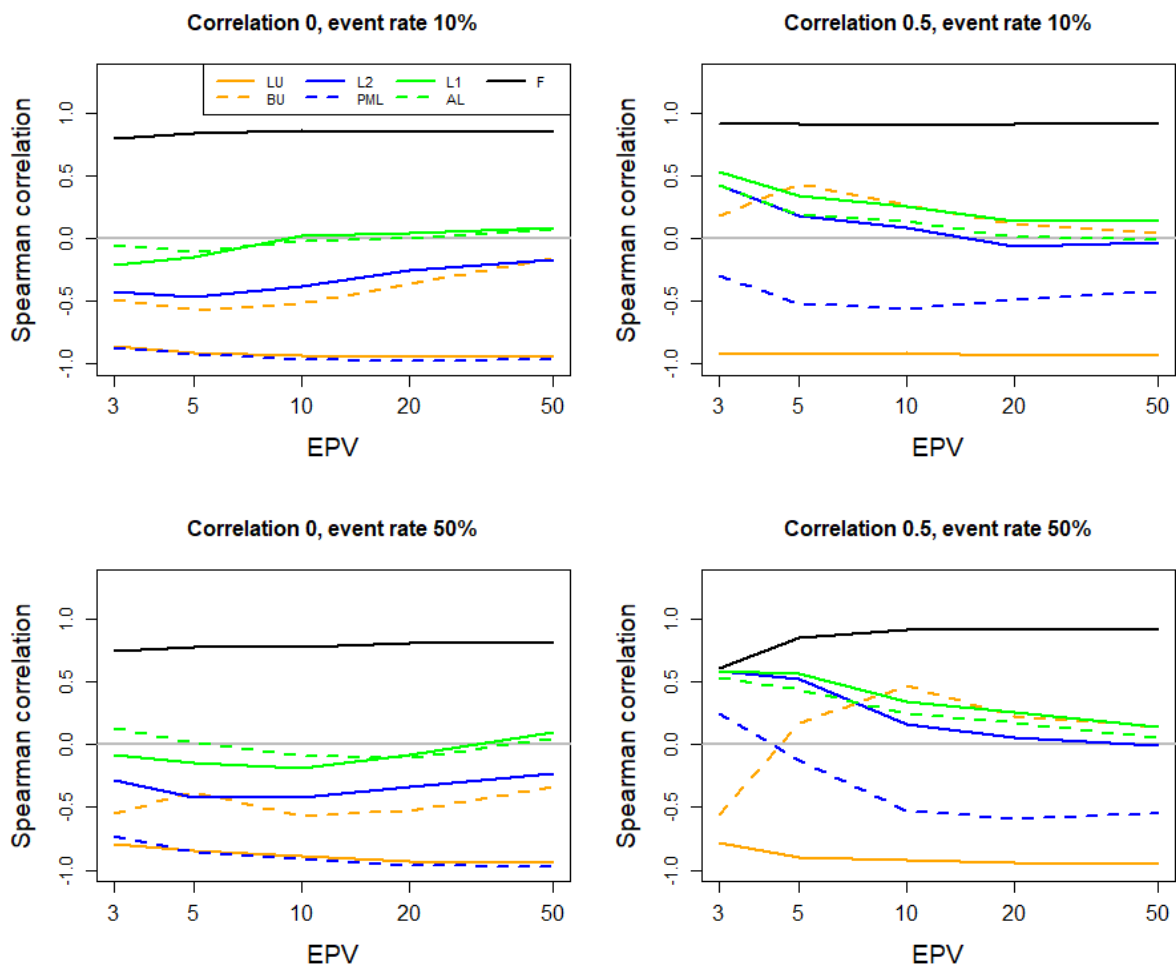


B. Scenarios with 5 true and 5 noise predictors



ew

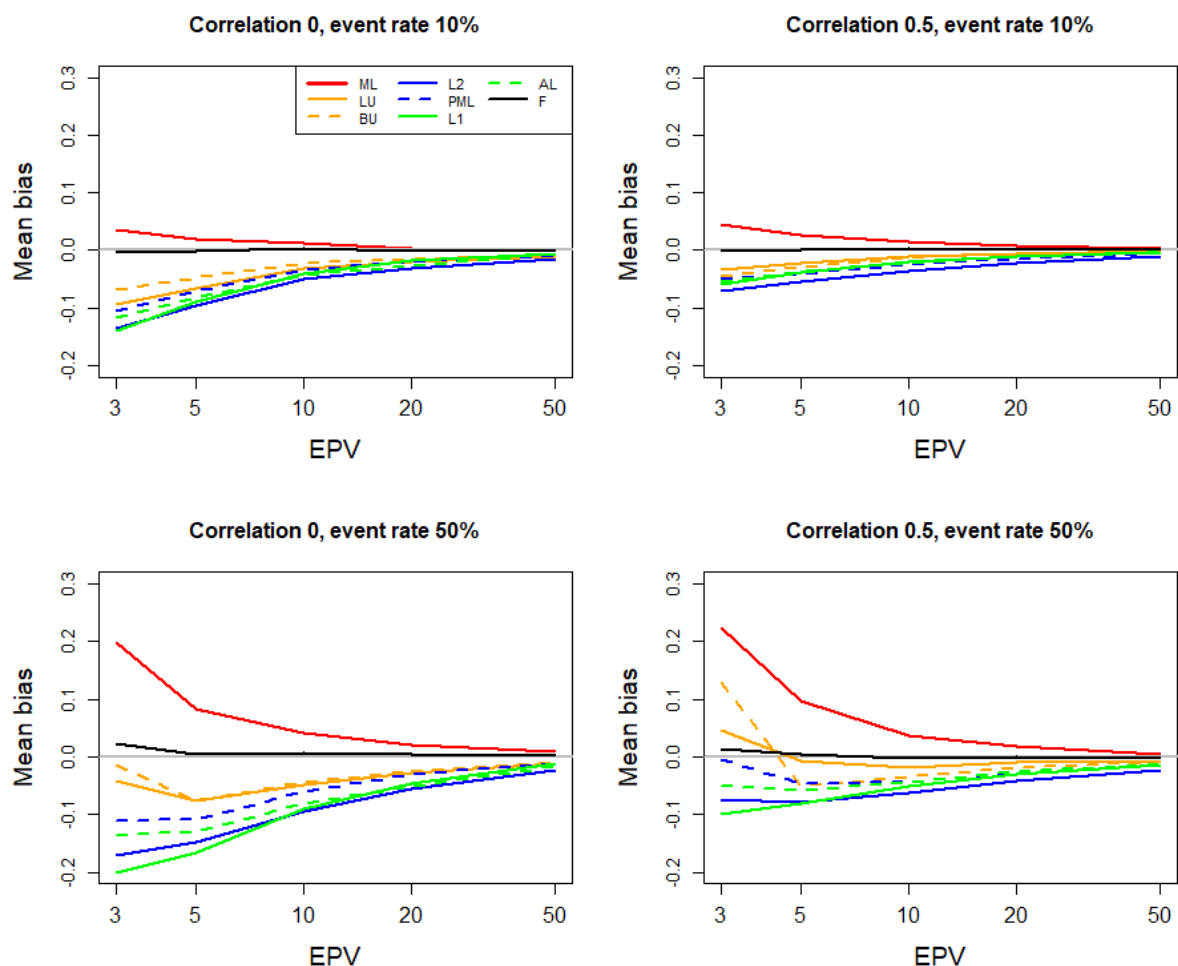
C. Scenarios with 10 true predictors



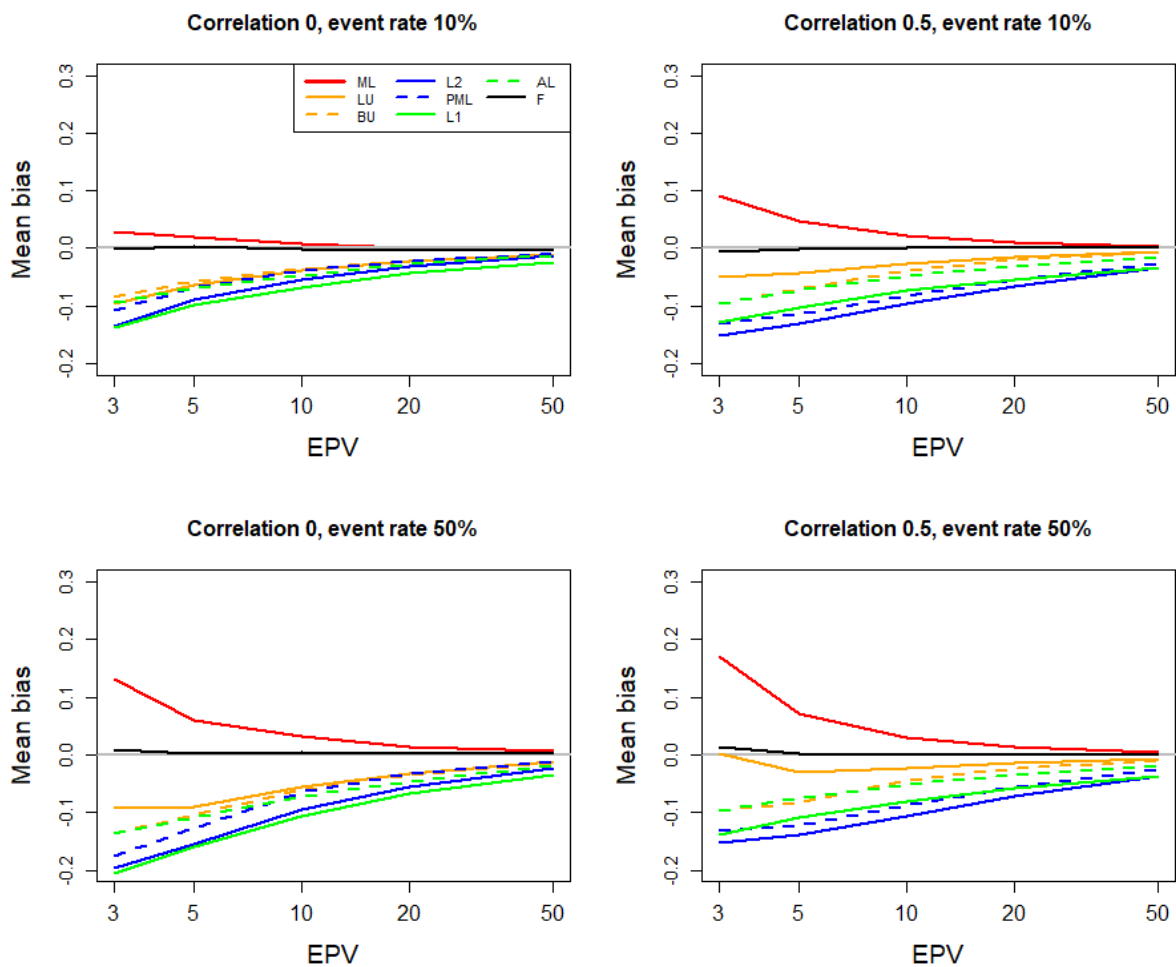
new

Figure S9. Mean bias in true coefficients per scenario and method. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

A. Scenarios with 5 true predictors

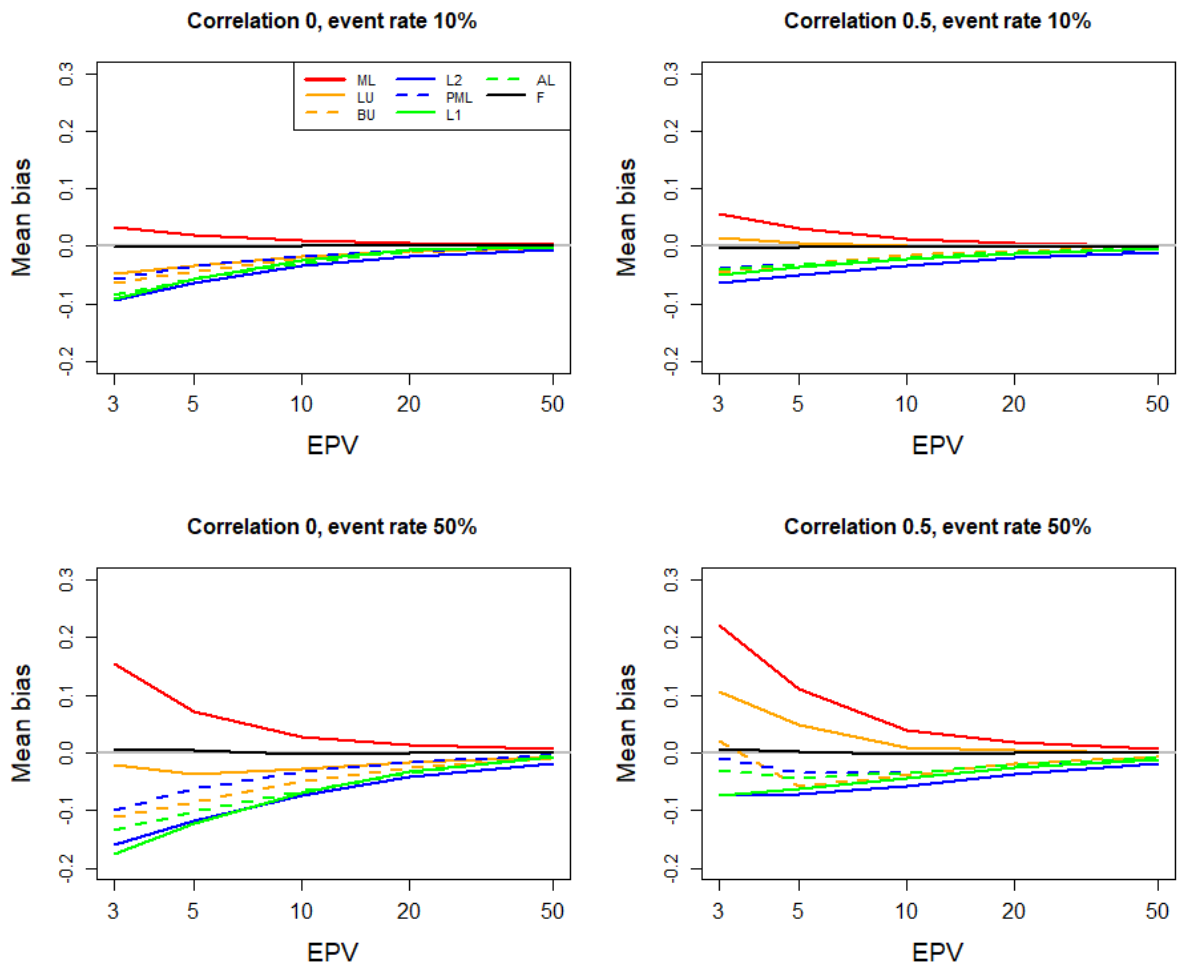


B. Scenarios with 5 true and 5 noise predictors



ew

C. Scenarios with 10 true predictors



new

Figure S10. Mean bias in noise coefficients per method, in scenarios with 5 true and 5 noise coefficients. ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on bootstrap; L2, classical ridge regression; PML, Harrell's penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F, Firth's correction.

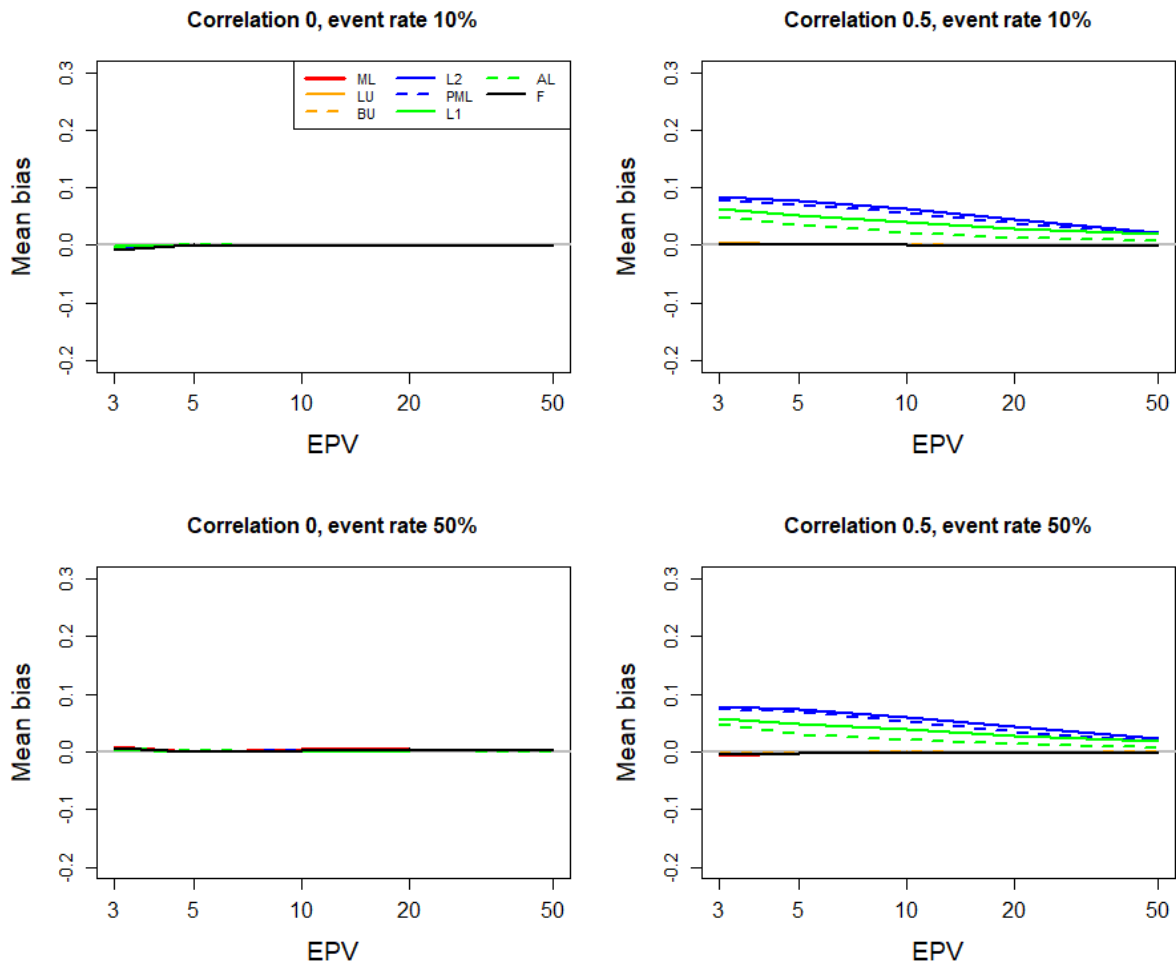
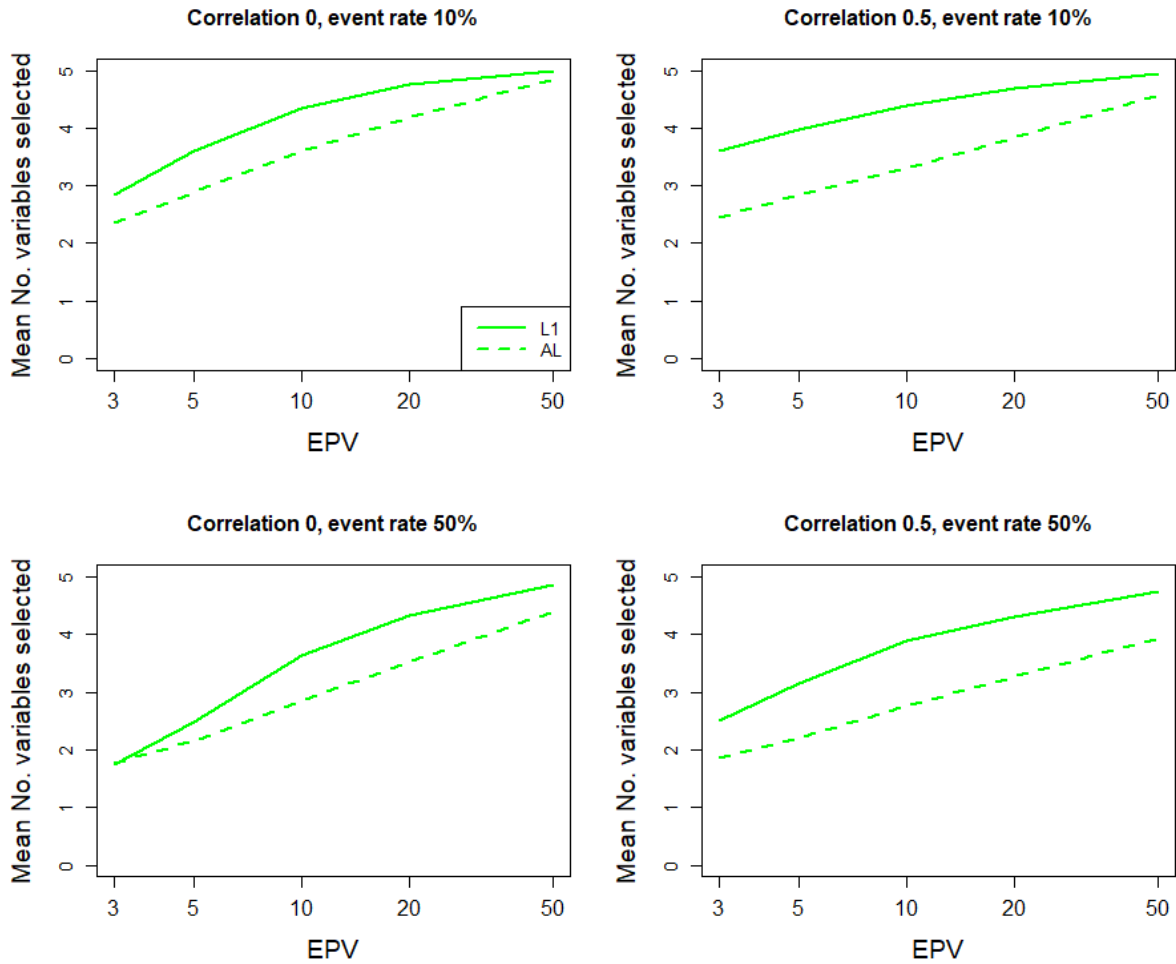
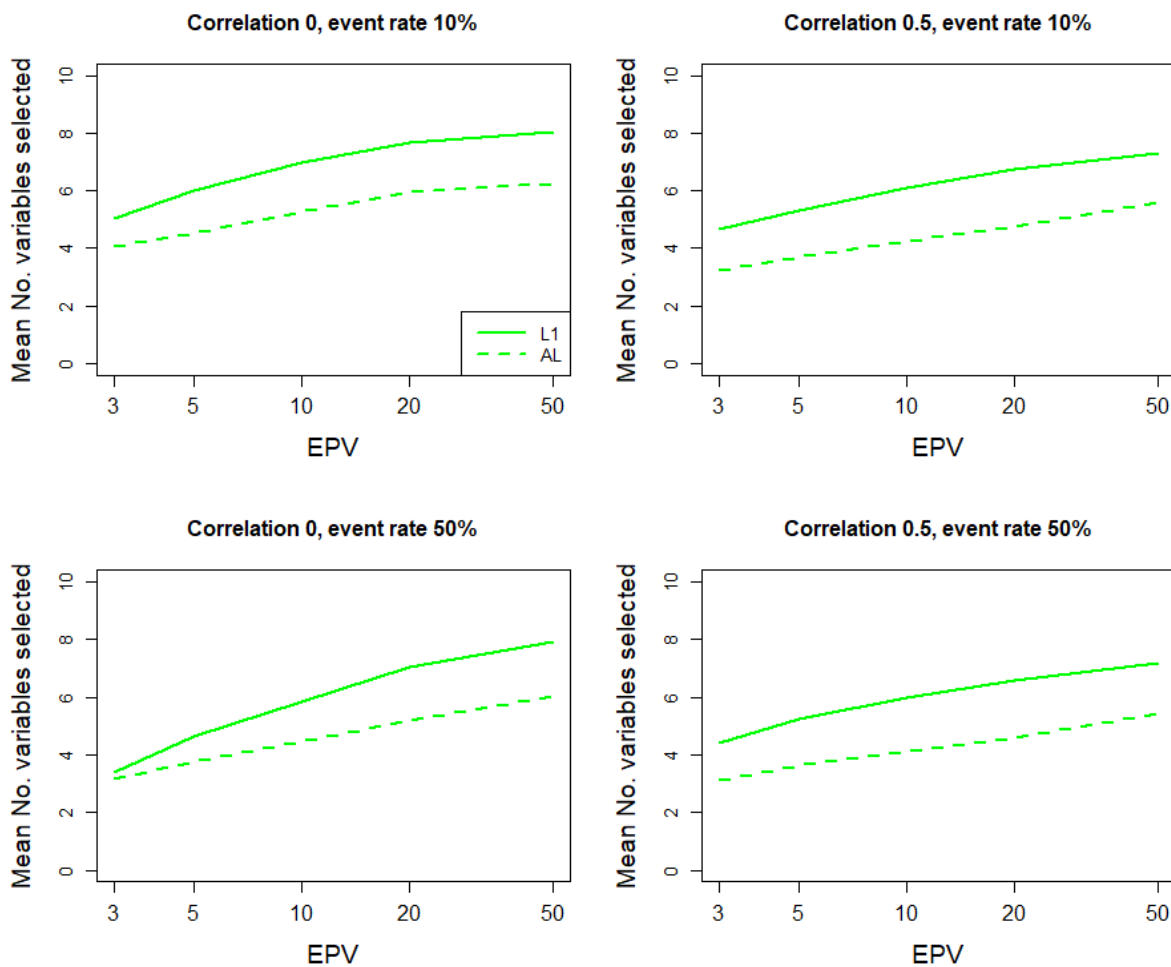


Figure S11. Mean number of selected variables per scenario, for methods based on LASSO procedures. L1, LASSO regression; AL, adaptive LASSO.

A. Scenarios with 5 true predictors

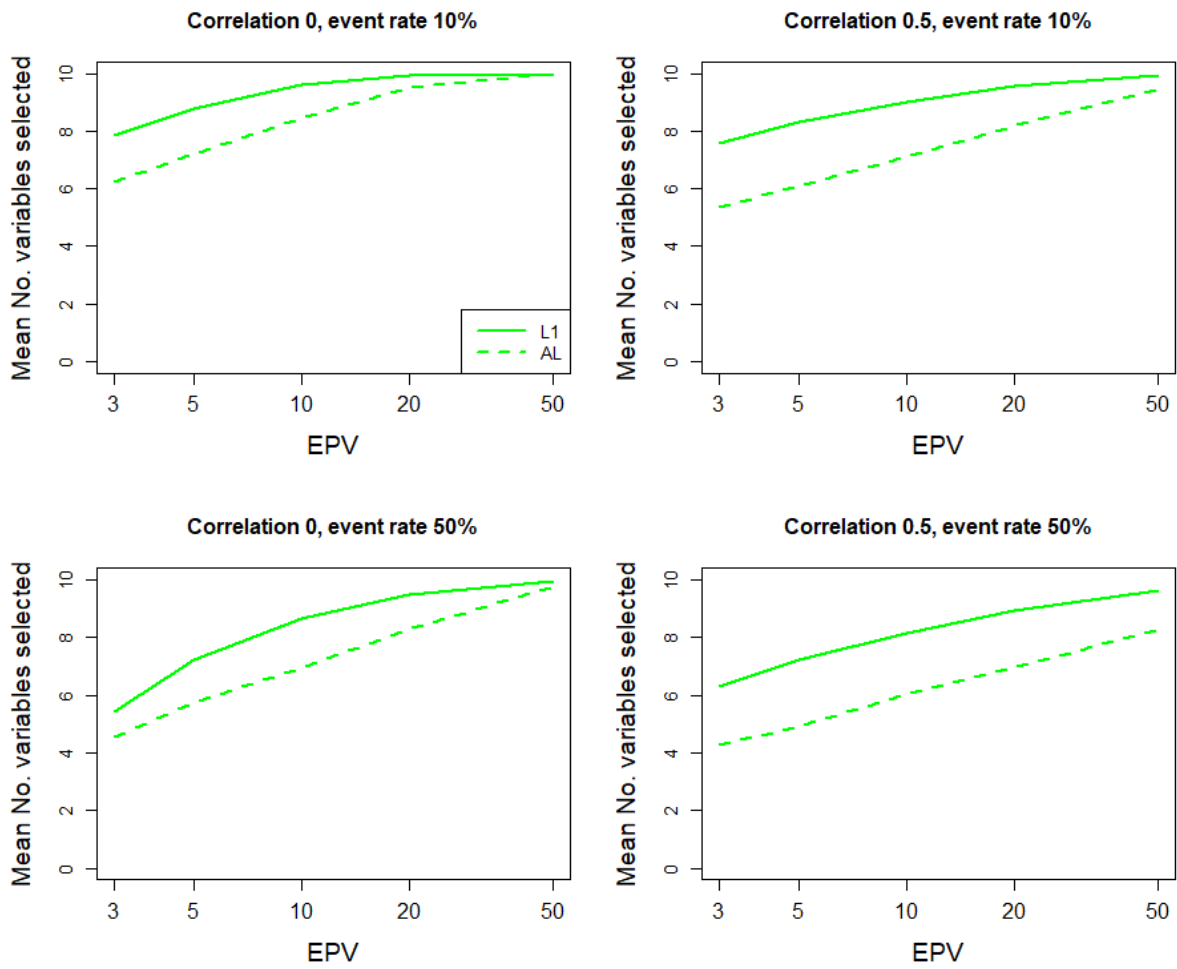


B. Scenarios with 5 true and 5 noise predictors



ew

C. Scenarios with 10 true predictors



new

Figure S12. Mean number of selected noise coefficients per method, in scenarios with 5 true and 5 noise coefficients. L1, LASSO regression; AL, adaptive LASSO.

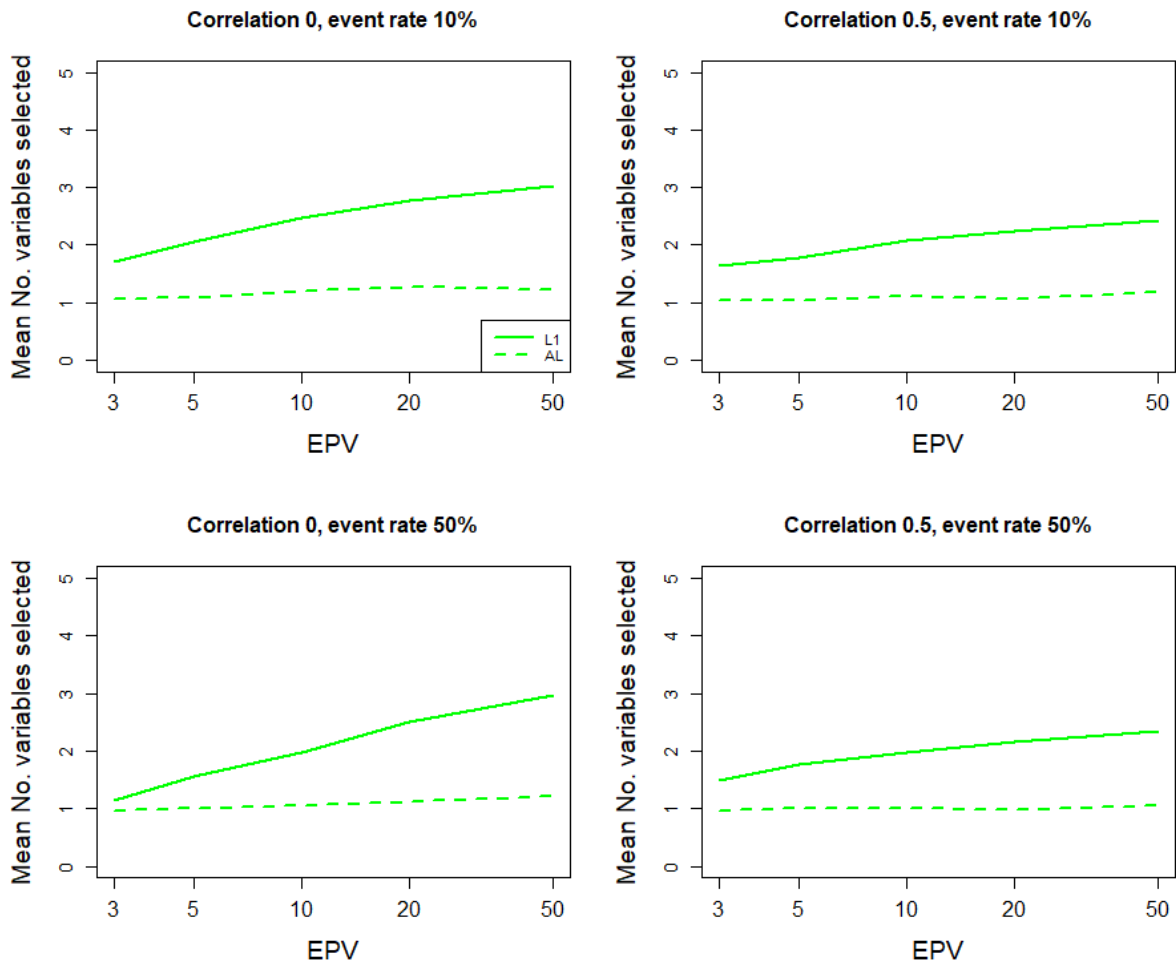


Table S5. Summary of simulation runs where LASSO or adaptive LASSO selected no variables at all. L1, LASSO (L1 penalty); AL, adaptive LASSO.

Simulation scenario	L1 runs, n (%)	AL runs, n (%)
3 EPV, 5 true predictors, 0 correlation, 0.5 event rate	334 (33%)	189 (19%)
5 EPV, 5 true predictors, 0 correlation, 0.5 event rate	172 (17%)	94 (9%)
3 EPV, 5 true + 5 noise predictors, 0 correlation, 0.5 event rate	165 (17%)	61 (6%)
3 EPV, 5 true predictors, 0 correlation, 0.1 event rate	121 (12%)	73 (7%)
3 EPV, 5 true predictors, 0.5 correlation, 0.5 event rate	102 (10%)	45 (5%)
3 EPV, 10 true predictors, 0 correlation, 0.5 event rate	69 (7%)	22 (2%)
5 EPV, 5 true + 5 noise predictors, 0 correlation, 0.5 event rate	33 (3%)	12 (1%)
5 EPV, 5 true predictors, 0 correlation, 0.1 event rate	26 (3%)	12 (1%)
5 EPV, 5 true predictors, 0.5 correlation, 0.5 event rate	22 (2%)	6 (1%)
3 EPV, 5 true + 5 noise predictors, 0 correlation, 0.1 event rate	20 (2%)	4 (<1%)
10 EPV, 5 true predictors, 0 correlation, 0.5 event rate	15 (2%)	12 (1%)
3 EPV, 5 true + 5 noise predictors, 0.5 correlation, 0.5 event rate	12 (1%)	4 (<1%)
5 EPV, 10 true predictors, 0 correlation, 0.5 event rate	6 (1%)	0
3 EPV, 5 true + 5 noise predictors, 0.5 correlation, 0.1 event rate	5 (1%)	0
3 EPV, 5 true predictors, 0.5 correlation, 0.1 event rate	3 (<1%)	1 (<1%)
3 EPV, 10 true predictors, 0 correlation, 0.1 event rate	1 (<1%)	0
10 EPV, 5 true + 5 noise predictors, 0 correlation, 0.5 event rate	1 (<1%)	0
20 EPV, 5 true predictors, 0 correlation, 0.5 event rate	1 (<1%)	0

1
2
3
4
5
6
7
8
9 **Regression shrinkage methods for clinical prediction models do not guarantee**
10 **improved performance: simulation study**
11
12
13
14
15
16

17 *Ben Van Calster*^{1,2}, *Maarten van Smeden*^{2,3}, *Bavo De Cock*^{1,4}, *Ewout W Steyerberg*²
18
19

20
21
22
23 ¹ KU Leuven, Department of Development and Regeneration, Herestraat 49 box 805, 3000 Leuven,
24 Belgium
25

26
27 ² Department of Biomedical Data Sciences, Leiden University Medical Center, PO Box 9600, 2300 RC
28 Leiden, Netherlands
29

30
31
32 ³ Department of Clinical Epidemiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA
33 Leiden, Netherlands
34

35
36 ⁴ KU Leuven, Department of Accountancy, Finance and Insurance, Naamsestraat 69 box 3525, 3000
37 Leuven, Belgium
38
39

40
41
42
43
44 E-mails ben.vancalster@kuleuven.be, M.van_Smeden@lumc.nl, E.W.Steyerberg@lumc.nl
45
46
47
48

49 Word count of main text: 4134
50
51
52
53
54
55
56
57
58
59
60

Corresponding author: Ben Van Calster, ben.vancalster@kuleuven.be, +32 16 377788

Abstract

When developing risk prediction models on datasets with limited sample size, shrinkage methods are recommended. Earlier studies showed that shrinkage results in better predictive performance on average. This simulation study aimed to investigate the variability of regression shrinkage on predictive performance for a binary outcome. We compared standard maximum likelihood with the following shrinkage methods: uniform shrinkage (likelihood-based and bootstrap-based), penalized maximum likelihood (ridge) methods, LASSO logistic regression, adaptive LASSO, and Firth's correction. In the simulation study, we varied the number of predictors and their strength, the correlation between predictors, the event rate of the outcome, and the events per variable. In terms of results, we focused on the calibration slope. The slope indicates whether risk predictions are too extreme ($\text{slope} < 1$) or not extreme enough ($\text{slope} > 1$). The results can be summarized into three main findings. First, shrinkage improved calibration slopes on average. Second, the between-sample variability of calibration slopes was often increased relative to maximum likelihood. In contrast to other shrinkage approaches, Firth's correction had a small shrinkage effect but showed low

1
2
3
4
5
6
7
8
9 variability. Third, the correlation between the estimated shrinkage and the optimal
10 shrinkage to remove overfitting was typically negative, with Firth's correction as the
11 exception. We conclude that, despite improved performance on average, shrinkage often
12 worked poorly in individual datasets, in particular when it was most needed. The results
13 imply that shrinkage methods do not solve problems associated with small sample size
14 or low number of events per variable.
15
16
17
18
19
20
21
22
23
24
25

26 **Keywords**

27
28
29
30
31
32 Clinical risk prediction models; Firth's correction; logistic regression; maximum
33 likelihood; penalized likelihood; shrinkage
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1. Introduction

When developing clinical prediction models, the ultimate aim is to obtain risk estimates that work well on patients that were not used to develop the model.¹ To do so, we have to keep statistical overfitting under control. Assuming that data collection was done carefully, and according to standardized procedures and definitions, the values in a dataset reflect (1) true underlying distributions of and associations between variables, and (2) some amount of random variability. Overfitting occurs when a prediction model also captures these random idiosyncrasies of the development dataset, which by definition do not generalize to new data from the same population.² The risk of an overfitted model increases when the model building strategy is too ambitious for the available data, for example when the number of variables that are tested as potential model predictors is large given the available sample size.

A well-known rule of thumb for sample size for prediction models is to have at least 10 events per variable (EPV).³⁻⁶ For binary outcomes, the number of events is the number of cases in the smallest of the two outcome levels. ‘Variables’ actually refers to the number of parameters that are considered for inclusion in the model (excluding intercepts). Some parameters may be checked but not included in the final model, and

1
2
3
4
5
6
7
8
9 variables may be modeled using more than one parameter. Recent research has
10 indicated that the $EPV \geq 10$ rule is too simplistic, and highlights that there are no good
11 rules of thumb regarding sample size.⁷⁻¹¹ Therefore, the use of shrinkage methods is
12 recommended when sample size is small.^{5,6} Several studies have suggested that model
13 performance improves on average when shrinkage methods are applied.^{5,9,12-17} Some
14 have suggested that shrinkage may be needed for EPV values up to 20 if the model is
15 prespecified.¹ When variable selection has to be performed to develop the model, the
16 required EPV for reliable selection may increase to 50.¹
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 Most regression shrinkage methods deliberately induce bias in the coefficient estimates,
32 by shrinking them towards zero, in order to reduce the expected variance in the
33 predictions. As a consequence, for models with a binary outcome, these methods aim to
34 prevent predicted risks that are too extreme, i.e. where small risks are underestimated,
35 and high risks overestimated. This leads to better expected mean squared error of the
36 predictions.¹⁸ Since prediction focuses on reliable predictions, inducing bias in the
37 model coefficients is not a key concern. Therefore, it seems that the use of shrinkage
38 methods is always good when sample size is limited. Moreover, standard maximum
39 likelihood estimation suffers from small sample bias leading to exaggerated coefficient
40 estimates (i.e. away from zero).^{6,19} However, some observations are puzzling. Hans van
41 Houwelingen already noted that 'it is surprising to observe that the estimated shrinkage
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 factors can be quite off the mark and are negatively correlated with optimal shrinkage
10 factor'.²⁰ This would imply that shrinkage methods shrink too little when it is really
11 needed, and vice versa. However, van Houwelingen's paper included only small
12 simulation study focusing on uniform shrinkage factors. It is of interest to see whether
13 this also occurs with other approaches to regression shrinkage, such as LASSO, ridge,
14 and Firth's correction.^{19,21,22} Other studies suggest that some methods result in too much
15 shrinkage on average, as indicated by an average calibration slope larger than
16 one.^{9,14,16,23} In Box 1, we present an illustration dealing with a prediction model for
17 ovarian cancer diagnosis,²⁴ to illustrate that standard regression and regression
18 shrinkage may be more variable in performance than many would think.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

36 The aim of this simulation study was to investigate the performance of various modern
37 shrinkage approaches for the validity of clinical prediction models that are developed
38 with small number of predictors relative to the total sample size (low dimensional). This
39 implies a situation in which some preselection of potentially important predictors has
40 been done before the modeling (e.g. by expert opinion or based on previous studies).
41
42 We address the performance on average, as well as performance for individual
43 simulation runs. The latter is done by evaluating the between-sample variability in the
44 amount of shrinkage provided by various methods, and the correlation between
45 estimated shrinkage and optimal shrinkage.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

2. Materials and methods

2.1. Regression methods

We will apply standard logistic regression based on maximum likelihood estimation, and compare this to a collection of shrinkage methods within the context of logistic regression. We apply likelihood-based and bootstrap-based uniform shrinkage methods,^{12,25} methods that directly shrink coefficient estimates without or with variable selection,^{21,22,26-29} and Firth's penalized likelihood.^{19,30} We will discuss each method in what follows.

Standard logistic regression. This is the reference method, in which coefficients are determined by maximum likelihood (ML). Hence, no shrinkage is applied here. When the outcome variable Y equals 1 for an event and 0 for a non-event, the probability of an event ($Y = 1$) for patient i (π_i) is estimated based on a weighted combination of p predictor variables X_j . We define π_i as $P(Y = 1|\mathbf{x}_i)$, with $i = 1, \dots, n$, and $\mathbf{x}_i =$

$(1, x_{1,i}, \dots, x_{p,i})'$. Assuming only linear effects and no interactions between the predictors, the logistic regression has the following form:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^p \beta_j x_{ij} = \mathbf{x}_i' \boldsymbol{\beta},$$

where $\pi_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}}$, and $\boldsymbol{\beta}$ a column vector containing the intercept α and the coefficients β_j . Coefficient estimates $\hat{\alpha}$ and $\hat{\beta}_j$ are obtained by finding the maximum of the log-likelihood function:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \{y_i \log(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \log(1 - \pi_i(\boldsymbol{\beta}))\}.$$

Likelihood-based uniform shrinkage (LU). This method uses the likelihood-ratio statistic to compute a uniform shrinkage factor

$$s_{LU} = \frac{\chi_{model}^2 - df}{\chi_{model}^2},$$

where χ_{model}^2 is the likelihood-ratio statistic of the fitted model based on standard maximum likelihood and df is the degrees of freedom for the number of candidate predictors considered for the model.²⁵ The shrunk model coefficients are then calculated as $\hat{\beta}_{j,LU} = s_{LU} \hat{\beta}_j$. After adjusting the coefficients, we re-estimated the intercept to guarantee that the average predicted risk equaled the event rate.

1
2
3
4
5
6
7
8
9
10
11
12 *Bootstrap-based uniform shrinkage (BU)*. The uniform shrinkage factor s can also be
13
14 computed using a bootstrap procedure:¹²
15
16
17
18
19

- 20 1. A bootstrap sample is taken from the original data sample, that is, a random
21 sample with replacement of the same size as the original sample.
 - 22 2. If a selection procedure was used to select variables this is also applied in the
23 bootstrap samples. The regression coefficients are estimated again on the bootstrap
24 sample, $\hat{\beta}_{bt}$.
 - 25 3. The linear predictor for each of the observations in the original sample is
26 calculated using $\hat{\beta}_{bt}$.
 - 27 4. In the original sample, the linear predictor obtained in the previous step is used to
28 predict the outcome using maximum likelihood. Retain the coefficient for the
29 regression of the linear predictor.
 - 30 5. Repeat the procedure, steps one to four, and the average coefficient from step four
31 provides the shrinkage factor s_{BU} . We used 200 repetitions.
 - 32 6. The shrunk coefficients are calculated as $\hat{\beta}_{j,BU} = s_{BU}\hat{\beta}_j$.
 - 33 7. Re-estimate the intercept using maximum likelihood while keeping $\hat{\beta}_{j,BU}$ fixed.
- 34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 *Classical ridge logistic regression.* Regression shrinkage is implemented via the ridge
10 penalty, also known as the quadratic or L2-penalty.²¹ Ridge regression was extended to
11 logistic regression initially by Schaefer and colleagues, and later by Le Cessie and Van
12 Houwelingen.^{26,27} The following penalized version of the log-likelihood function is
13 maximized:
14
15
16
17
18
19

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p \beta_j^2.$$

20
21
22
23
24 The tuning parameter, λ , controls the amount of shrinkage. Ridge regression shrinks the
25 estimated coefficients towards zero on average, with higher values of λ leading to more
26 shrinkage. This implicitly induces bias in the coefficients. Note that coefficients will not
27 be shrunk to zero and that the intercept term is not penalized. The shrinkage parameter λ
28 is a hyperparameter that has to be estimated ('tuned'). We used 10-fold cross-validation
29 to find the value for λ that minimized the deviance, using a grid of 251 possible values
30 between zero (no shrinkage) and 64 (very large shrinkage). The 250 non-null values
31 were equidistant on logarithmic scale. We used the glmnet R package to implement
32 ridge logistic regression.³²
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

49 *General penalized maximum likelihood estimation.* Ridge logistic regression is a special
50 case of penalized maximum likelihood (PML) that maximizes the following function:²⁸
51
52
53
54
55
56
57
58
59
60

$$\ell(\boldsymbol{\beta}) - 0.5\lambda \sum_{j=1}^p (s_j \beta_j)^2,$$

where s_j are scaling factors that allow more flexibility than classical ridge. In our study, we will apply the method as suggested by Harrell.²⁸ We set the scale factors to the standard deviation of the predictor. As our predictors are simulated as standard normal variable, and we standardize the variables before fitting models, this approach does not differ from classical ridge. However, Harrell suggests to tune the shrinkage parameter based on a Akaike Information Criterion instead of cross-validation, because it is faster and performs slightly better.²⁸ Following Harrell's suggestion, the tuning parameter was chosen using the corrected Akaike Information Criterion using a similar grid as for classical ridge.^{28,33} The rms R package was used to implement this method. In tables and figures, we refer to this method with the abbreviation PML, and to classical ridge regression with the abbreviation L2.

Classical LASSO logistic regression. LASSO is similar to ridge, but uses the L1-penalty that poses a constraint on the sum of the absolute value of the estimated coefficients.²²

For logistic regression, the LASSO optimizes the following function:

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p |\beta_j|.$$

1
2
3
4
5
6
7
8
9 The L1-penalty allows coefficients to be shrunk to zero, and hence LASSO performs
10 variables selection as well. The shrinkage parameter was tuned using cross-validation in
11 the same way as for classical ridge logistic regression. The glmnet R package was used.
12
13
14
15
16
17
18

19 *Adaptive LASSO (AL)*. The Adaptive LASSO is a variant of the LASSO where a weight
20 is given for each parameter in the penalty term, in order to obtain variable selection
21 consistency.²⁹ The optimized function is:
22
23
24
25

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p w_j |\beta_j|,$$

26
27 where $w_j = \frac{1}{|\hat{\beta}_j^{init}|^\gamma}$ contains adaptive weights. The $\hat{\beta}_j^{init}$ are initial coefficient estimates
28 for the predictors. We used the maximum likelihood estimate $\hat{\beta}_j$ as $\hat{\beta}_j^{init}$, and fixed γ at
29 unity.^{15,29} Adaptive LASSO shrinks higher absolute values of $\hat{\beta}_j^{init}$ less than lower
30 values. We tuned the shrinkage parameter using cross-validation as for classical
31 LASSO. The glmnet R package was used.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47 *Firth's penalized likelihood*. Firth developed a procedure to remove the first order bias
48 in the regression coefficients based on maximum likelihood.^{19,30} To do so, modified
49 score functions are used to estimate model coefficients. This avoids problems with
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 separation, but also shrinks the coefficients. In addition, Firth's correction reduces the
10 variance. In terms of the log-likelihood, Firth's correction optimizes

$$\ell(\boldsymbol{\beta}) + 0.5\log|I(\boldsymbol{\beta})|,$$

16
17 where $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. We used the `logistf` R
18 package to implement this method. For making predictions based on Firth's correction,
19 the intercept has to be corrected.¹⁶ We used the same intercept re-estimation procedure
20 as for the uniform shrinkage methods.
21
22
23
24
25
26
27
28
29

30 2.2. Simulation setup

31
32
33
34
35
36 We simulated data to predict a binary outcome. We used a full factorial simulation setup
37 varying the following factors: EPV, the number and strength of predictors, the
38 correlation between predictors, and the outcome event rate (Table 1). In total, this gave
39 us 60 simulation scenarios. In the setting with five true predictors, the true coefficients
40 of the predictors were 0.2, 0.2, 0.2, 0.5, and 0.8. These values were based on the
41 Cohen's d measure of effect size, and would correspond to having three weak predictors
42 (odds ratio 1.22), one moderate predictor (odds ratio 1.65), and one strong predictor
43 (odds ratio 2.23).³¹ In the setting with 10 true predictors, six had a coefficient of 0.2,
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 two had a coefficient of 0.5 and two had a coefficient of 0.8. Noise predictors had
10 coefficients of 0. The chosen values of the simulation factors had an impact on the true
11 c-statistic (i.e. area under the receiver operating characteristic curve) of the model, the
12 sample size of the simulated datasets, and the number of cases with an event (Table 2).
13
14
15
16
17
18
19
20

21
22 For every scenario, the simulations were performed as follows. First, for each of 1 000
23 000 individuals the predictor values were generated by draws from a standard
24 multivariate normal distribution, with equal pairwise correlations. The true model
25 formula (linear predictor) was applied to each patient, with the intercept chosen to
26 obtain the target event rate (Table 2). The inverse logit of the linear predictor was the
27 true risk for that individual. Then, the outcome for each patient was generated through a
28 Bernoulli trial using the true risk. A different dataset, but also with 1 000 000
29 individuals, was generated for model validation. Predictors and outcomes were
30 generated analogous to the development population, which means that our out-of-
31 sample performance corresponds to a large sample internal validation setting.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 We executed 1 000 simulation runs per simulation condition. For each run, we
49 generated a development dataset of the appropriate size (Table 2) by randomly drawing
50 without replacement from the development population. The event rate was fixed at the
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 target value in each development dataset by applying stratified sampling. Next, the
10 predictor variables were standardized, and all types of models were fitted. When using
11 standard maximum likelihood, separation was suggested when R warned for fitted
12 probabilities of zero or one, or when the model did not converge. In these
13 circumstances, results for standard maximum likelihood were replaced with results
14 based on Firth's correction, because this is a situation where the use of the method is
15 indicated.^{19,30} For LU and BU, the shrinkage factor s was calculated for the model using
16 Firth's correction (with bootstrap models for BU also based on Firth's correction).
17 Harrell's suggested PML method often resulted in an error when there was the
18 suggestion of separation. In these cases, we used Firth's correction instead of the PML
19 algorithm. In this way, we could avoid the exclusion of samples that were suggestive of
20 separation.³⁴ For logistic regression with bootstrap uniform shrinkage, bootstrap models
21 suggestive of separation were replaced by other bootstrap replicates without separation.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 The resulting models were validated on the accompanying full validation dataset. We
44 calculated the c-statistic and the calibration slope. Because the development and
45 validation data are based on identical populations, the calibration intercept was of little
46 interest and therefore not calculated.³⁵ At internal validation (i.e. when the underlying
47 population is the same), the calibration slope measures bias of risk predictions in terms
48 of spread.^{35,36} A slope below unity suggests that predictions are too extreme: low risks
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 are underestimated, high risks are overestimated. A slope above unity suggests the
10 opposite. We calculated median slopes to assess the deviation from the target value of
11 unity. To investigate the variability in the slope, we calculated the median absolute
12 deviation (MAD) of the $\log(\text{slope})$. To combine bias (deviation of slope from unity on
13 average), and variability, we calculated root mean squared distance from the target
14 value (RMSD) of the $\log(\text{slope})$ over the 1 000 runs. We used the logarithm of the slope
15 to acknowledge its asymmetry. A slope of 0.5 (half the target) corresponds to a similar
16 quantitative deviation to a slope of two (double the target), but in opposite directions.
17 The RMSD was calculated as the square root of the mean of $(\log(1) - \log(\text{slope}))^2$
18 over the 1 000 runs. Finally, we calculated the Spearman correlation between the
19 estimated shrinkage and the optimal shrinkage over the 1 000 simulation runs. The
20 optimal shrinkage was defined as $\log(1) - \log(\text{slope}_{\text{ML}})$, with slope_{ML} the slope for
21 the standard maximum likelihood model. The estimated shrinkage for a specific
22 shrinkage approach was defined as $\log(\text{slope}_{\text{shrinkage}}) - \log(\text{slope}_{\text{ML}})$. To calculate
23 MAD, RMSD, and correlations, we winsorized slopes at 0.01 to avoid problems with
24 rare instances of negative calibration slopes. When no variables were selected by
25 (adaptive) LASSO, the calibration slope was arbitrarily set at 10 to reflect the extreme
26 amount of underfitting.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

R code used for the simulations can be found at GitHub

(<https://github.com/benvancalster/shrinkagesim/>).

3. Results

There were few runs where separation was suggested (Table S1), except in the scenario with three EPV, 10 true predictors, 0.5 correlation and 0.5 event rate. Generally, results differed little between the five predictor and 10 prediction scenarios, therefore we focus here on the scenarios with five true predictors for the main document. Detailed results for all scenarios are provided in supplementary tables and figures.

3.1. Performance on average

The median calibration slope approached unity for all methods as EPV increased (Figure 1, Figure S1, Table S2). The standard maximum likelihood model yielded the lowest median calibration slopes. For classical ridge regression, the median slope at lower EPV values was consistently above unity, suggesting too much shrinkage on average. Harrell's PML and LASSO were better, but in many scenarios showed median

1
2
3
4
5
6
7
8
9 slopes above unity as well. Other methods generally had median slopes below unity,
10 with bootstrap uniform shrinkage usually having median slopes closest to unity. The use
11 of Firth's correction was slightly better than maximum likelihood.
12
13
14
15
16
17
18

19 The average c-statistics also converged to their respective true values as EPV increased
20 (Figure S2). By design, uniform shrinkage had the same c-statistics as regular maximum
21 likelihood. When predictors were correlated, classical ridge and Harrell's PML had
22 highest c-statistics. When predictors were uncorrelated and no noise predictors were
23 present, LASSO had lower c-statistics than the maximum likelihood model. Adaptive
24 LASSO only had better discrimination than maximum likelihood when noise predictors
25 were present. Firth's correction did not improve the c-statistic.
26
27
28
29
30
31
32
33
34
35
36
37
38

39 3.2. Variability in the applied shrinkage 40 41 42 43 44

45 For the scenarios with five true predictors, pairwise correlations of 0.5 between
46 predictors, and an event rate of 50%, box plots of the calibration slopes over the 1 000
47 simulation runs are shown in Figure 2. For all scenarios, box plots are given in Figure
48 S3, and MAD in Figure S4. The variability of the calibration slope after shrinkage was
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 larger than the variability based on maximum likelihood, except when Firth's correction
10 was used. Firth's correction consistently reduced variability (Figure S4). This increased
11 variability was particularly strong when EPV is low, and correlations between
12 predictors were low. Only when there were 10 true predictors with high
13 intercorrelations, most shrinkage methods had lower variability than maximum
14 likelihood.
15
16
17
18
19
20
21
22
23
24
25

26 Generally, shrinkage methods improved the RMSD relative to the maximum likelihood
27 model (Table S3, Figure 3, Figure S5). However, LASSO, adaptive LASSO, classical
28 ridge and Harrell's PML often had higher RMSD than maximum likelihood when
29 predictors were uncorrelated and EPV or sample size was low. Classical ridge and
30 Harrell's PML often showed higher RMSD than other methods when predictors were
31 correlated and EPV was high. Two methods, the bootstrap uniform shrinkage and
32 Firth's correction, always had lower RMSD than maximum likelihood.
33
34
35
36
37
38
39
40
41
42
43
44
45

46 Box plots of the c-statistics also showed high between-sample variability for all
47 methods (Figure S6).
48
49
50
51
52
53
54
55
56
57
58
59
60

3.3. Correlation between estimated and optimal shrinkage

The Spearman correlation between estimated and optimal shrinkage was typically negative (Figure 4, Table S4, Figures S7-8). Firth's correction was the exception with consistently positive correlations. LASSO-based methods typically had the lowest negative correlations (closest to zero). For these methods, correlations were highest, and in particular cases even positive, in settings with more highly correlated predictors. The highest positive correlations between estimated and optimal shrinkage were found when there were 10 true predictors, there was non-zero true correlation between the predictors and the EPV was low.

3.4. Results for coefficient estimates and variable selection

Coefficient estimates of true predictors were exaggerated when the maximum likelihood model was used (Figure S9). The bias decreased with increasing EPV. Using Firth's correction removed the bias. All other shrinkage methods induced negative bias and consistently underestimated the coefficients. With respect to noise predictors, classical

1
2
3
4
5
6
7
8
9 ridge, Harrell's PML, LASSO, and adaptive LASSO had positive bias in the estimated
10 coefficients when there was correlation between predictors (Figure S10).
11
12
13
14
15
16

17 Regarding variable selection, adaptive LASSO selected less predictors than standard
18 LASSO implementations (Figure S11). In simulation scenarios with noise predictors,
19 these predictors were selected more often with increasing EPV, except when adaptive
20 LASSO was used (Figure S12). Table S5 summarizes how often these methods selected
21 no variables at all.
22
23
24
25
26
27
28
29
30
31

32 4. Discussion 33 34 35 36 37

38 In this paper, we assessed the performance of various shrinkage methods for clinical
39 risk prediction models using simulations. Our key results were the following. First,
40 shrinkage led to calibration slopes that were on average closer to the ideal value of unity
41 than maximum likelihood. Firth's correction improved the slope least among the
42 considered methods. Classical ridge, and to a lesser extent Harrell's PML and LASSO,
43 tended to shrink too much overall. Second, the performance of the shrinkage methods
44 was highly variable, especially when sample size was relatively low. The exception was
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Firth's correction, which showed remarkably stable performance. Despite the increased
10 variance, the RMSD of the calibration slopes was usually lower for shrinkage methods
11 compared to standard maximum likelihood. This was notably the case for Firth's
12 correction, due to its limited variability, but also for bootstrap uniform shrinkage. Third,
13 we commonly observed that the estimated shrinkage was inversely correlated with the
14 optimal shrinkage. This corroborated the early observation by van Houwelingen,²⁰ and
15 implies that shrinkage often does least when it is needed most. Firth's correction was
16 again the exception, with consistently positive correlations. Fourth, there were
17 differences between the shrinkage methods. A key parameter to this end is the RMSD,
18 because it combines bias in and variability of the calibration slope. Based on RMSD,
19 Firth's correction and bootstrap uniform shrinkage would be the preferred methods.
20 Shrinkage using the bootstrap uniform shrinkage factor performed remarkably well,
21 perhaps because this method explicitly uses the calibration slope for shrinkage
22 estimation. Firth's penalized likelihood almost surely improved performance over
23 maximum likelihood, with low variability and positive correlation with optimal
24 shrinkage. Important advantages of Firth's correction that lead to its stability are that it
25 does not require the estimation of a tuning parameter, and that it shrinks extreme risk
26 estimates. However, the magnitude of shrinkage was small.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 These results have implications. Although shrinkage works on average by bringing the
10 calibration slope closer to unity, it may not work as anticipated for any given dataset.
11
12 The variability in the estimated shrinkage was particularly high when sample size was
13
14 low. Thus, the use of shrinkage does not justify using lower sample size for the
15
16 development of prediction models. When sample size is low, it may even be advisable
17
18 not to build a prediction model. Alternatively, a less complicated model can be
19
20 considered, for example by discarding many predictors a priori. In a previous study in
21
22 the context of survival prediction models,¹⁴ the authors suggested that it may be
23
24 possible to develop an acceptable model with EPV of 2.5 if methods like ridge or
25
26 LASSO are used, although acknowledged that more work was required.¹⁴ We cannot
27
28 defend this suggestion based on our results.
29
30
31
32
33
34
35
36
37

38 We have to be careful about recommendations with respect to specific shrinkage
39
40 approaches, because the study was not designed to inform fully on their relative merits.
41
42 For example, classical ridge with tuning based on 10-fold cross-validation led to poorer
43
44 median calibration slopes than Harrell's PML estimation with tuning based on the
45
46 corrected Akaike Information Criterion, but had less variability in the calibration slope
47
48 (Figure S4). More research should study the impact of specific combinations of
49
50 shrinkage and tuning methods.
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12 The first limitation of our study was the focus on low-dimensional settings for which
13 predictors were largely pre-specified. It would be relevant to investigate the issues of
14 high variability and negative correlation in high-dimensional settings, settings where
15 both sample size and the number of potential predictors are large (such as in some
16 electronic health record studies). Second, we focused on normally distributed predictors,
17 although typical applications also contain non-normal predictors such as skewed
18 continuous predictors or categorical predictors. However, this does not invalidate the
19 key results of our paper. We anticipate the performance variability to become even be
20 larger when a mixture of predictor types are used. Third, we deliberately fixed event
21 rates in simulated datasets, because in prediction model applications one has to go by
22 the event rate that is observed in the data at hand. A downside of this choice is that our
23 results ignore sampling variability in the event rate in observational cross-sectional or
24 cohort studies. Such variability in the observed event rate may further worsen variability
25 in performance. Finally, we investigated many well-known shrinkage methods.
26
27 Nevertheless, it may be interesting to investigate whether our findings can be confirmed
28 in other approaches, such as elastic net, smoothly clipped absolute deviation (SCAD),
29 weighted fusion, or machine learning methods.³⁷⁻³⁹
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Our results are in line with previous work. In line with a large recent simulation study,
10 model performance in our study was related to event rate even when EPV was fixed.⁹
11 Further, the results are consistent with the recommendation to base sample size on a
12 maximal expected level of shrinkage.¹⁰ In accordance with earlier work, we observed
13 that methods like ridge or LASSO may have the tendency to shrink too much on
14 average.^{8,14-16} Perhaps the use of cross-validation may contribute to this, because
15 shrinkage parameter tuning is based on datasets with reduced sample size. However, in
16 contrast with earlier claims,^{6,14} the bootstrap uniform shrinkage method performed
17 relatively well in our simulations. These claims were based on simulations with 2.5
18 EPV, which is lower than the values considered in our study. Our results do not support
19 the development of prediction models with such low EPV with any method, although
20 more work on settings with very low event rates may be of interest.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 In conclusion, shrinkage improves performance on average. The larger variability in
41 calibration slope with the use of shrinkage methods, and the negative correlation
42 between estimated and optimal shrinkage, suggest that shrinkage may not work well for
43 any given dataset. Firth's correction is a notable exception, with reduced variability and
44 a positive correlation between estimated and optimal shrinkage. However, the amount
45 of shrinkage it applied was modest. Overall, the use of shrinkage is not a solution to the
46 problem of low sample size or low EPV. In such cases, more fundamental changes are
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 needed, such as refraining from the development of a model, increasing sample size, or
10
11 reducing a priori the number of predictors if this is clinically acceptable.
12
13
14
15
16

17 **Declarations of Conflicting Interests**

18
19
20
21
22

23 The Authors declare that there is no conflict of interest.
24
25
26
27
28

29 **Funding**

30
31
32
33
34

35 The author(s) disclosed receipt of the following financial support for the research,
36 authorship, and/or publication of this article: This work was supported by the Research
37 Foundation – Flanders (FWO) [grant number G0B4716N]; and the Internal Funds KU
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

58 **ORCID iD**

59
60
Ben Van Calster 0000-0003-1613-7450

1
2
3
4
5
6
7
8
9 Maarten van Smeden 0000-0002-5529-1541

10
11
12 Bavo De Cock 0000-0002-1310-6336

13
14
15 Ewout W Steyerberg 0000-0002-7787-0122

16
17
18
19
20
21 **References**

- 22
23
24
25
26
27 1. Steyerberg EW. *Clinical prediction models* (2nd edition). New York: Springer,
28 2019.
29
30
31 2. Babyak MA. What you see may not be what you get: a brief, nontechnical
32 introduction to overfitting in regression-type models. *Psychosom Med* 2004; 66:
33 411-421.
34
35
36 3. Harrell FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved
37 prognostic prediction. *Stat Med* 1984; 3: 143-152.
38
39
40 4. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events
41 per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373-1379.
42
43
44 5. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with
45 logistic regression analysis: a comparison of selection and estimation methods in
46 small data sets. *Stat Med* 2000; 19: 1059-1079.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 6. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk
10 prediction model when there are few events. *BMJ* 2015; 351: h3868.
- 11
12
13 7. Courvoisier DS, Combescure C, Agoritsas T, et al. Performance of logistic
14 regression modeling: beyond the number of events per variable, the role of data
15 structure. *J Clin Epidemiol* 2011; 64: 993-1000.
- 16
17
18 8. van Smeden M, de Groot JA, Moons KG, et al. Sample size for binary logistic
19 prediction models: Beyond events per variable criteria No rationale for 1 variable
20 per 10 events criterion for binary logistic regression analysis. *BMC Med Res*
21 *Methodol* 2016; 16: 163.
- 22
23
24 9. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic
25 prediction models: Beyond events per variable criteria. *Stat Meth Med Res* 2019;
26 28: 2455-2474.
- 27
28
29 10. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a
30 multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat*
31 *Med* 2019; 38: 1276-1296.
- 32
33
34 11. Ogundimu EO, Altman DG and Collins GS. Adequate sample size for developing
35 prediction models is not simply related to events per variable. *J Clin Epidemiol*
36 2016; 76: 175-182.
- 37
38
39 12. Steyerberg EW, Eijkemans MJC and Habbema JDF. Application of shrinkage
40 techniques in logistic regression analysis: a case study. *Stat Neerl* 2001; 55: 76-88.
- 41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 13. Steyerberg EW, Eijkemans MJC, Harrell FE Jr, et al. Prognostic modeling with
10
11 logistic regression analysis: in search of a sensible strategy in small data sets. *Med*
12
13 *Decis Making* 2001; 21: 45-56.
- 14
15 14. Ambler G, Seaman S and Omar RZ. An evaluation of penalised survival methods
16
17 for developing prognostic models with rare events. *Stat Med* 2012; 31: 1150-1161.
- 18
19 15. Pavlou M, Ambler G, Seaman SR, et al. Review and evaluation of penalised
20
21 regression methods for risk prediction in low-dimensional data with few events.
22
23 *Stat Med* 2016; 35: 1159-1177.
- 24
25 16. Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events:
26
27 accurate effect estimates and predictions? *Stat Med* 2017; 36: 2302-2317.
- 28
29 17. De Jong VMT, Eijkemans MJC, Van Calster B, et al. Sample size considerations
30
31 and predictive performance of multinomial logistic prediction models. *Stat Med*
32
33 2019; 38: 1601-1619.
- 34
35 18. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B* 1983; 45: 311-354.
- 36
37 19. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; 80: 27-
38
39 38.
- 40
41 20. van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve
42
43 predictive accuracy. *Stat Neerl* 2001; 55: 17-34.
- 44
45 21. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal
46
47 problems. *Technometrics* 1970; 12: 55-67.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 22. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B*
10 1996; 58: 267-288.
11
12
13 23. Musoro JZ, Zwinderman AH, Puhan MA, et al. Validation of prediction models
14 based on lasso regression with multiply imputed data. *BMC Med Res Methodol*
15 2014; 14: 116.
16
17
18 24. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish
19 between the benign and malignant adnexal mass before surgery: a multicenter study
20 by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; 23: 8794-
21 8801.
22
23
24 25. Van Houwelingen JC and le Cessie S. Predictive value of statistical models. *Stat*
25 *Med* 1990; 9: 1303-1325.
26
27
28 26. Schaefer RL, Roi LD and Wolfe RA. A ridge logistic estimator. *Commun Stat*
29 *Theory Methods* 1984; 13: 99-113.
30
31
32 27. Le Cessie S and van Houwelingen JC. Ridge estimators in logistic regression. *J R*
33 *Stat Soc C* 1992; 41: 191-201.
34
35
36 28. Harrell FE Jr. *Regression modeling strategies* (2nd edition). New York: Springer,
37 2015.
38
39
40 29. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; 101:
41 1418-1429.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9 30. Heinze G and Schemper M. A solution to the problem of separation in logistic
10 regression. *Stat Med* 2002; 21: 2409-2419.
- 11
12
13 31. Pencina MJ, D'Agostino RB Sr, Pencina KM, et al. Interpreting incremental value
14 of markers added to risk prediction models. *Am J Epidemiol* 2012; 176: 473-481.
- 15
16
17 32. Friedman JH, Hastie T and Tibshirani R. Regularization paths for generalized linear
18 models via coordinate descent. *J Stat Softw* 2010; 33: 1.
- 19
20
21 33. Hurvich CM and Tsai CL. Regression and time series model selection in small
22 samples. *Biometrika* 1989; 76: 297-307.
- 23
24
25 34. Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic
26 regression: causes, consequences, and control. *Am J Epidemiol* 2018;187:864-870.
- 27
28
29 35. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk
30 models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; 74: 167-
31 176.
- 32
33
34 36. Cox DR. Two further applications of a model for binary regression. *Biometrika*
35 1958; 45: 562-565.
- 36
37
38 37. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R*
39 *Stat Soc B* 2005; 67: 301-320.
- 40
41
42 38. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its
43 oracle properties. *J Am Stat Assoc* 2001; 96: 1348-1360.
- 44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

39. Daye ZJ and Jeng XJ. Shrinkage and model selection with correlated variables via weighted fusion. *Comput Stat Data Anal* 2009; 53: 1284-1298.

For Peer Review

1
2
3
4
5
6
7
8
9 Figure 1. Median calibration slopes for the scenarios with five true predictors.

10
11
12 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
13 shrinkage based on bootstrapping; L2, classical ridge regression; PML, Harrell's
14 penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F,
15 logistic regression with Firth's correction.
16
17
18
19

20
21
22
23
24
25 Figure 2. Box plots of the calibration slope over the 1 000 simulation runs for scenarios
26 with five true predictors, no correlation between predictors, and 50% event rate. The
27 events per variable is indicated in the top left. The numbers at the bottom are the root
28 mean squared distances (RMSD) of the log of the calibration slopes. The length of the
29 whiskers is at most 1.5 times the interquartile range. Calibration slopes are winsorized
30 at 0.1 and 10 for visualization purposes.
31
32
33
34
35
36
37
38

39 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
40 shrinkage based on bootstrapping; L2, classical ridge regression; PML, Harrell's
41 penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F,
42 logistic regression with Firth's correction.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 Figure 3. Root mean squared distance (RMSD) of the logarithm of the calibration slope
10 over 1 000 simulation runs for scenarios with five true predictors.
11

12
13
14 ML, maximum likelihood; LU, uniform shrinkage based on likelihood; BU, uniform
15 shrinkage based on bootstrapping; L2, classical ridge regression; PML, Harrell's
16 penalized maximum likelihood; L1, LASSO regression; AL, adaptive LASSO; F,
17 logistic regression with Firth's correction.
18
19
20
21
22

23
24
25
26
27 Figure 4. Scatter plots of the slope after shrinkage versus the slope based on maximum
28 likelihood (no shrinkage) for the scenario with five true predictors, no correlation
29 between predictors, 50% event rate, and three events per variable. Each point represents
30 one of the 1 000 simulation runs. The blue line is the diagonal, where both slopes are
31 the same. The green lines show the ideal slope (unity). Red circles refer to simulation
32 runs where maximum likelihood resulted in a slope above unity.
33
34
35
36
37
38
39

40
41 LU, uniform shrinkage based on likelihood; BU, uniform shrinkage based on
42 bootstrapping; L2, classical ridge regression; PML, Harrell's penalized maximum
43 likelihood; L1, LASSO regression; AL, adaptive LASSO; F, logistic regression with
44 Firth's correction.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Box 1. Clinical illustration: prediction model for ovarian cancer diagnosis.

In 2005, the International Ovarian Tumor Analysis group published its first ultrasound-based risk prediction models to diagnose ovarian malignancy in patients that are selected for surgery.²⁴ The dataset of 1066 patients were randomly split in a development part of 754 (191 with a malignancy) and a validation part of 312 (75 with a malignancy). For the model, over 40 predictors were considered, totaling 52 parameters. The EPV was 3.7 (191/52). Data-driven variable selection was used in the context of standard logistic regression (no shrinkage), leading to a model with 12 predictors. Using the dataset from this study, the model had a calibration intercept of 0.007 and a calibration slope of 1.09 on the validation part. Contrary to expectation, the observed slope suggested mild underfitting: the estimated risks were too close to the overall outcome prevalence. If likelihood-based uniform shrinkage factor were used,²⁵ predictors coefficients would have been multiplied by 0.89. This implies a shrinkage of 11%, which seems little given the data-driven selection among 52 parameters in a dataset of moderate size. With this method, the calibration slope on the validation part would have been 1.22. Hence, shrinkage worsened the calibration of the model. Obviously, the small size of the validation part set implied considerable random variation. Nevertheless, this illustrates that a thorough assessment of the variability of standard logistic regression and alternatives based on shrinkage is important.

Table 1. Overview of the simulation factors in the full factorial simulation design.

Simulation factor	Factor levels
Events per variable	3, 5, 10, 20, 50
Predictors	five true predictors; 10 true predictors; five true and five noise predictors
Correlation between predictors	0, 0.5
Outcome event rate	0.1, 0.5

Table 2. Overview of the characteristics of the 60 simulation scenarios

Predictors	Correlation	Event rate	EPV	Events	Sample size	True c statistic	Model intercept		
Five true predictors, or five true + five noise predictors	0	0.1	3	15	150	0.75	-2.57		
			5	25	250				
			10	50	500				
			20	100	1000				
			50	250	2500				
	0.5	0.5	3	15	30	0.74	0		
			5	25	50				
			10	50	100				
			20	100	200				
			50	250	500				
	0.5	0.1	0.1	3	15	150	0.83	-2.98	
				5	25	250			
				10	50	500			
				20	100	1000			
50				250	2500				
0.5		0.5	0.5	3	15	30	0.81	0	
				5	25	50			
				10	50	100			
				20	100	200			
				50	250	500			
10 true predictors	0	0.1	3	30	300	0.82	-2.88		
			5	50	500				
			10	100	1000				
			20	200	2000				
			50	500	5000				
		0.5	0.5	0.5	3	30	60	0.80	0
					5	50	100		
					10	100	200		
					20	200	400		
					50	500	1000		
	0.5	0.1	0.1	3	30	300	0.93	-4.34	
				5	50	500			
				10	100	1000			
				20	200	2000			

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

			50	500	5000		
		0.5	3	30	60	0.91	0
			5	50	100		
			10	100	200		
			20	200	400		
			50	500	1000		

For Peer Review