Journal of Clinical Epidemiology 199 (2026) 112423

ORIGINAL RESEARCH

# A preregistered simulation study provided evidence on the appropriate use of data-driven variable selection in descriptive multivariable regression modeling

Theresa Ullmann[a], Georg Heinze[a], Michael Kammer[a,b], Daniela Dunkler[a,*], for TG2 of the STRATOS Initiative

[a]Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria
[b]Division of Nephrology and Dialysis, Department of Medicine III, Medical University of Vienna, Vienna, Austria

Accepted 14 July 2026; Published online 23 July 2026

## Abstract

**Objectives:** Multivariable regression modeling is central in epidemiological research. Common practice often uses data-driven variable selection to identify the variables most strongly associated with the outcome (the goal of descriptive modeling), but this may lead to omission of relevant covariates, inclusion of noise variables, biased coefficient estimates, and unstable models. Although methodological recommendations exist, guidance based on systematic and neutral comparison studies remains limited. Therefore, we aimed to provide evidence-based guidance on the appropriate use of data-driven variable selection in epidemiological regression modeling, with a focus on descriptive modeling.

**Study Design and Setting:** We conducted a comprehensive simulation study comparing commonly used variable selection approaches, including backward elimination (BE) with various selection criteria and variants of the least absolute shrinkage and selection operator (Lasso). Methods were evaluated across realistic epidemiological scenarios varying in outcome type (binary or continuous), sample size, and signal-to-noise ratio. Performance was assessed using measures relevant to descriptive modeling: variable selection rates (TPR: true positive rates and FPR: false positive rates), model selection rates, and bias of regression coefficients.

**Results:** No single selection method performs uniformly well across all settings. Acceptable performance requires large sample sizes and/or strong signal-to-noise ratios. Under such favorable conditions, BE with a mild selection criterion ($\alpha = 0.5$) performs best in terms of high TPR, BE with stricter criteria ($\alpha = 0.05$ or Bayesian information criterion) performs best in terms of low FPR, and BE with Akaike information criterion strikes a balance. Lasso methods do not achieve optimal performance when evaluated against descriptive modeling criteria.

**Conclusion:** Variable selection methods may complement substantive knowledge−driven specification but only under favorable conditions and not as the default strategy in epidemiological analyses. The choice of method should be guided by the desired trade-off between high TPR and low FPR, for which we provide a practical decision framework. 

*Keywords:* Backward elimination; Penalized estimation; Lasso; Multivariable modeling; Regression modeling; Simulation

## 1. Introduction

In observational epidemiology, multivariable regression models are central tools for estimating associations between an outcome and multiple independent variables. After identifying an initial variable set using subject-matter knowledge and/or data availability, researchers frequently apply data-driven variable selection post hoc to simplify models or construct adjustment sets, often without explicit consideration of the underlying analytic objective or the conditions under which such procedures are defensible.

Funding: This work was supported by the Austrian Science Fund (FWF) [I-4739-B to DD]. For open access purposes, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

* Corresponding author. Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Spitalgasse 23, Wien 1090, Austria.

*E-mail address:* daniela.dunkler@meduniwien.ac.at (D. Dunkler).

https://doi.org/10.1016/j.jclinepi.2026.112423
0895-4356/

**Plain Language Summary**

Researchers often use statistical models to study how different variables, such as age, gender, or medical conditions, are related to health outcomes. To develop these models, decisions have to be made about which variables to include and which to exclude. In practice, this is commonly done applying a data-driven variable selection method on the observed data. However, this approach can produce a misleading final statistical model. In this study, we examined how well popular data-driven variable selection methods are able to distinguish variables that are truly associated with the outcome from irrelevant noise variables. We used computer simulations that mimic real-world epidemiological data to compare different approaches, including commonly used techniques such as BE and Lasso. We evaluated these methods under a range of conditions, including small versus large sample sizes and strong versus weak relationships between variables and outcomes. As expected, we found that no method works well in all situations. In particular, all variable selection methods tended to perform poorly when the dataset was small or when the relationship between the variables and the outcome was weak. In these cases, important variables may be missed, irrelevant ones may be included, estimated effects can be biased, and the selected model can vary considerably depending on the specific data used. When conditions are more favorable (eg, with large datasets and strong relationships), variable selection methods can perform reasonably well. Simpler approaches such as BE can strike a reasonable balance between keeping important variables and excluding irrelevant ones, depending on how strict or lenient the statistical criteria for removing variables are. More complex methods, such as Lasso, did not consistently perform better. Overall, our results suggest that data-driven variable selection should not be used automatically. Instead, researchers should first rely on existing scientific knowledge to decide which variables to include. If data-driven variable selection is used subsequently, it should only be done under appropriate conditions and with a clear understanding of its limitations. With a decision framework, we provide practical guidance to help researchers decide when and how to use these methods more appropriately.

Popular selection methods include backward elimination (BE), which starts with a full model (FU) containing all covariates from the initial set and iteratively removes variables based on a significance threshold $\alpha$ or an information criterion. Penalized estimation methods (eg, Lasso [1]) are another option. Reviews consistently show that selection is common in applied epidemiology and medicine [2−7], despite long-standing concerns [8−11] regarding bias, instability, and overinterpretation. These concerns include omitting important covariates, selecting noise variables, biasing coefficients, and producing unstable selections.

Applied epidemiologists and methodological collaborators would benefit from clearer guidance [12] on when to use data-driven selection in analyses, which methods to choose under common conditions, and how to select method-specific parameters such as $\alpha$.

Some recommendations already exist. Harrell [8] generally advises against selection but notes that BE with lenient $\alpha$ (0.5) may remove irrelevant variables. Heinze et al [10] provide guidance using events-per-variable (EPV), defined as the number of events (logistic regression) or the sample size $n$ (linear regression) divided by the number of covariates in FU. They recommend avoiding selection when EPV $\leq$ 10, exercising caution for EPV of 10−25 due to instability and coefficient bias, and limiting expectations to more reliable yet still imperfect performance when EPV $>$25. These recommendations conclude by calling for systematic comparisons of commonly used methods to inform applied practice.

To address this gap, we conducted a neutral simulation comparing popular selection methods in common epidemiological scenarios [13], with the explicit aim of translating empirical results into practical guidance for routine analyses. A central feature is the explicit consideration of the modeling objective [14−17]—either descriptive, predictive, or explanatory—which critically determines whether and how selection should be applied. We focus on descriptive modeling, where the objective is to characterize associations in a parsimonious and interpretable way rather than to predict outcomes or estimate causal effects. Specifically, descriptive modelling may aim either to estimate associations in a pre-specified model, or to identify variables most strongly associated with the outcome, for which data-driven selection may be considered (as in this paper); the term is sometimes used more narrowly to refer only to the pre-specified estimation aspect. We evaluate performance measures relevant to this objective: variable selection rates (VSRs), model selection rates (MSRs), and coefficient bias.

We restrict attention to variants of BE and the Lasso [18−21], as these are among the most commonly applied methods. Other methods—forward selection and univariable selection—are discussed in Section 4. We focus on logistic and linear regression with low-dimensional data, that is, the number of variables is lower than $n$. Finally, we synthesize our results into a practical decision framework intended to support transparent, informed choices about whether and how data-driven variable selection should be used in epidemiological practice.

The article is structured as follows: Section 2 summarizes the simulation design, Section 3 presents results, Section 4 discusses recommendations, and Section 5 concludes.

**What is new?**

**Key findings**

- Performance of data-driven variable selection methods depends strongly on events-per-variable and signal-to-noise ratios, with substantial deterioration under sparse data or weak signals.
- When the modeling goal is descriptive, backward elimination more reliably balances retention of relevant variables against exclusion of noise variables than Lasso-type procedures.

**What this study adds to what is known?**

- This study demonstrates the inherent trade-off between retaining true predictors and excluding noise variables and provides a practical framework for choosing selection methods accordingly.

**What is the implication and what should change now?**

- Data-driven variable selection should not be routine practice and should be used only under favorable conditions regarding events-per-variable and signal strength.
- Subject-matter—informed prespecification of candidate variables remains essential regardless of whether data-driven selection is subsequently applied.

## 2. Simulation design

We describe the simulation study using the ADEMP structure (aims, data-generating mechanisms, estimands and targets, methods, and performance measures) [22]. Details are provided in the peer-reviewed study protocol [13].

### 2.1. Aims

We examined the performance of selection methods in multivariable logistic and linear regression under varying

- sample size/EPV and
- population signal strength (population $R^2$).

### 2.2. Data-generating mechanisms

#### 2.2.1. Predictors and noise

To ensure a realistic data structure [23], correlations (Fig S1) and variable distributions were based on National Health and Nutrition Examination Survey [24]. Variables were binary, normal, or skewed (Fig S2) and generated using the normal-to-anything (NORTA) technique [25,26], which preserves marginal distributions and correlations. We considered 20 variables: 10 true predictors (“predictors”) and 10 noise variables (“noise”).

Predictor effects were assigned using standardized coefficients:

$$\left(\beta_1^{sd}, \ldots, \beta_{10}^{sd}\right)^t = (1.5, -1, 1, 0.75, 0.5, 0.5, 0.5, -0.5, -0.25, -0.25)$$

Predictors were ordered by effect size, with predictor 1 having the largest effect. Standardized effects were converted to raw coefficients using empirical standard deviations from a large simulated dataset ($n = 100,000$). Noise variables had a true effect of zero.

#### 2.2.2. The outcome Y

Continuous outcomes were generated by adding random noise to the linear predictor defined through assignment of effects to predictors, with noise chosen to achieve different signals:

- Strong signal: $R^2 = 0.45$
- Weak signal: $R^2 = 0.15$

For binary outcomes, a logistic model generated probabilities from the linear predictor, and outcomes were drawn from a Bernoulli distribution. Model parameters were adjusted to vary event rates and Cox-Snell $R^2_{CS}$, yielding corresponding population-level area under the receiver operating characteristic curve (AUC):

- Event rate 0.3:
  - Strong signal: $R^2_{CS} = 0.40$, AUC = 0.90
  - Weak signal: $R^2_{CS} = 0.13$, AUC = 0.73
- Event rate 0.05:
  - Strong signal: $R^2_{CS} = 0.16$, AUC = 0.94
  - Weak signal: $R^2_{CS} = 0.05$, AUC = 0.78

#### 2.2.3. Sample sizes n

To facilitate comparison between linear and logistic regression, we defined $n$ for linear regression and calculated corresponding $n$ for logistic regression with event rate 0.3 to achieve similar coefficient precision at each aligned $n$ (eg, at $n = 183$ in logistic regression, SEs match those at $n = 100$ in linear regression; Table 1). We also added $n$ for EPV = 25, as it was mentioned in recommendations [10].

For logistic regression with event rate 0.05, where precision alignment was unstable, we matched EPV with linear regression up to EPV = 25, yielding $n = 2000$ (EPV = 5), 4000 (EPV = 10), 8000 (EPV = 20), and 10, 000 (EPV = 25).

**Table 1.** Sample sizes $n$ and events-per-variable (EPV) for linear regression and logistic regression with event rate 0.3

| Regression | Label for *n:* | Very small | Small | Medium | Moderately large | Large | Very large | Extremely large | Moderately large |
|---|---|---|---|---|---|---|---|---|---|
| Logistic | $n$ | 183 | 365 | 730 | 1461 | 2922 | 5844 | 11,687 | *1667* |
| | EPV | 2.75 | 5.48 | 10.95 | 21.92 | 43.83 | 87.66 | 175.31 | *25* |
| Linear | $n$ | 100 | 200 | 400 | 800 | 1600 | 3200 | 6400 | *500* |
| | EPV | 5 | 10 | 20 | 40 | 80 | 160 | 320 | *25* |

### 2.3. Estimands and targets

We evaluated how accurately methods identified predictors and noise variables and how well true coefficients were estimated.

### 2.4. Methods

We assessed popular selection methods (details in the Supplementary Material A):

- BE(0.05), BE(0.50), BE(AIC), BE(BIC): BE with $\alpha = 0.05$, $\alpha = 0.5$, Akaike information criterion (AIC), or Bayesian information criterion (BIC)
- Lasso variants
  - Lasso(CV): Lasso [1] with 10-fold cross-validation (CV)
  - RLasso(CV), RLasso(BIC): relaxed Lasso [19,20] with 10-fold CV or BIC
  - AdaLasso(CV): adaptive Lasso [21] with 10-fold CV.
- FU: model containing all 20 variables, as reference.

For BE, information criterion selection corresponds to significance testing: AIC to $\alpha \approx 0.157$ and BIC to an $\alpha$ decreasing with increasing $n$ (eg, $\alpha \approx 0.032$ at $n = 100$, $\alpha \approx 0.002$ at $n = 10,000$). In this simulation, BIC represents the strictest BE criterion. For Lasso variants with CV, prediction performance on test folds was evaluated using deviance (logistic regression) or mean squared error (linear regression).

### 2.5. Performance measures

We evaluated performance by how well methods identified true predictors and estimated coefficients. The performance measures are defined in Table 2. The results were averaged over 2000 simulation repetitions.

## 3. Results

### 3.1. Variable selection rates: true positive rates and false positive rates

*Logistic regression with event rate 0.3*: Figure 1 shows VSR for each of the 10 predictors and 10 noise variables in the strong signal scenario (left column) and the weak signal scenario (right column), for BE methods (upper panel) and Lasso variants (lower panel).

The correlations of the variables influence their selection rates. For example, the true positive rate (TPR) of predictor 2 tends to be lower than that of predictor 3, although both predictors have the same effect size. This is because predictor 2 has a much higher multiple correlation, making it harder for selection methods to separate signal from noise.

For BE, higher significance thresholds increase TPR, but also false positive rate (FPR). By definition, FPRs correspond to selection thresholds (eg, FPR of BE(0.50) are $\approx$ 0.5).

In the strong signal scenario, mild BE (0.5) reaches TPR near 1 for all predictors at moderately large $n$, albeit with high FPR. Stricter criteria (0.05, AIC, BIC) reduce VSR, requiring large $n$ to reach TPR near 1.

The Lasso yields high VSR for predictors and noise variables, with FPR increasing with growing $n$. Relaxed Lasso yields lower VSR with BIC than with CV. For both tuning approaches, very large $n$ are required to reach a TPR of 1. Both relaxed Lasso variants struggle with correct (non) selection of variables with high multiple correlation. At smaller $n$, relaxed Lasso shows low TPR for predictor 2 (the second-largest predictor), which is highly correlated with others. One noise variable with high correlations is frequently selected; its FPR even increases with $n$. The adaptive Lasso attains TPR close to 1 for all predictors for large $n$, earlier than the relaxed Lasso. Its FPRs start relatively high; starting from large $n$ they begin to decline.

In the weak signal scenario, FPR resemble those in the strong signal scenario, but TPR are lower; larger $n$ are required to approach TPR near 1, even for BE(0.50) and the Lasso.

*Logistic regression with event rate 0.05* (Fig S3): Results resemble those for event rate 0.3 at corresponding EPV, but the low event rate requires much larger $n$ to reach the respective EPV. We simulated data only up to $n = 10,000$ (EPV = 25). Thus, even in the strong signal scenario, TPRs for some predictors remain below 1 with BE(0.05), BE(BIC), or either relaxed Lasso variant.

*Linear regression* (Fig S4): Results are very similar to logistic regression with event rate 0.3 when compared at $n$ aligned by coefficient precision. The aligned $n$ correspond to *smaller* $n$ but *higher* EPV in linear regression (Table 1), so smaller $n$ are sufficient (but higher EPV are necessary) to achieve comparable performance.

**Table 2.** Key performance measures for multivariable regression modeling with a descriptive goal

| Performance measure | Explanation |
|---|---|
| **Variable selection rates (VSR):**<br>True positive rate (TPR) for predictors,<br>False positive rate (FPR) for noise variables | **Answers which question?**<br>How likely is it for a variable to be selected?<br>**Approximation by simulation:**<br>Frequency with which a specific variable was selected across the simulation repetitions.<br>**Range & ideal values**:<br>0 to 1. Ideally, TPR should be 1, FPR should be 0.<br>In most practical situations, ideal TPR and FPR are unlikely. |
| **Model selection rates (MSR):**<br>MSR of true model,<br>MSR of any "overselection" model,<br>MSR of any "underselection model" | **Answers which question?**<br>How likely is it for a certain type of model to be selected?<br>• The *true model* consists exactly of the true predictors.<br>• Any *"overselection" model* includes all predictors plus at least one noise variable.<br>• Any *"under-selection" model* misses at least one predictor.<br>The definitions of under-selection and overselection models could also be adapted according to the application of interest [27].<br>**Approximation by simulation:**<br>Frequency with which a certain type of model was selected across the simulation repetitions.<br>**Range & ideal values:**<br>0 to 1. Ideally, a variable selection method should have a true model selection rate of 1, which is unlikely in most practical scenarios. |
| **Bias of regression coefficients:**<br>Conditional bias,<br>Unconditional bias | **Answers which question?**<br>• *Conditional bias*: If we apply a selection method, what is the expected bias of the regression coefficient of a certain variable, *given that this variable was actually selected*?<br>• *Unconditional bias*: If we apply a selection method, what is the expected bias of the regression coefficient of a specific variable, *regardless of whether the variable was actually selected*?<br>Unconditional bias results from a combination of (1) how often the variable is expected to be selected (selection rate) and (2) what the expected bias is if the variable was selected (conditional bias).<br>**Approximation by simulation:**<br>Bias is calculated using standardized coefficients for comparability across variables.<br>• Conditional bias: bias is averaged only across simulation repetitions where the variable was selected.<br>• Unconditional bias: bias is averaged across all simulation repetitions.In each repetition, an estimated coefficient of zero is plugged in for unselected variables.<br>**Range & ideal values**:<br>$-\infty$ to $\infty$. Ideally, bias should be 0. |

### 3.2. Model selection rates

*Logistic regression with event rate 0.3* (Fig 2): In the strong signal scenario, only BE(0.05) and BE(BIC) achieve moderate-to-high true model MSR for larger $n$. Among these, only BE(BIC) reaches a true model MSR near 1, but only at very large $n$. BE(0.05) achieves a true model MSR of ~0.6 for large $n$. As stable selection of the true model requires very large $n$, a pragmatic alternative is selecting an overselection model. Methods that consistently include all important variables—while accepting some irrelevant ones—are preferable up to moderately large $n$. This holds for BE(0.50) and the Lasso, which yield overselection MSR above 0.8 for medium $n$, and near 1 for moderately large $n$. Consequently, for medium or moderately large $n$, BE(0.50) or the Lasso may be preferable to BE(BIC), which exhibits high underselection MSR. For very small or small $n$, all methods underselect heavily, making FU—always an overselection model—the safest choice.

As expected, results deteriorate in the weak signal scenario. Even at extremely large $n$, the true model is recovered in only ~55% of repetitions for BE(0.05) and ~60% for BE(BIC). Compared to the strong signal scenario, larger $n$ are required to achieve high overselection MSR: BE(0.50) and the Lasso approach 1 only at very large $n$.

*Logistic regression with event rate 0.05* (Fig S5): Findings match those for event rate 0.3 at corresponding EPV. With data simulated only up to $n$ = 10, 000 (EPV = 25), true model MSR remain far below 1 even at the largest $n$. In the strong signal scenario, BE(0.05) and BE(BIC) reach true model MSR only slightly above 0.5 at $n$ = 10, 000. In the weak signal scenario, true model MSR are (near) zero across all methods and $n$.

*Linear regression* (Fig S6): Results resemble logistic regression with event rate 0.3 at aligned $n$. However, in the weak signal scenario, even BE(BIC) and BE(0.05) fail to identify the true model in at least 50% of repetitions, even at the largest $n$.

In many applied studies, $n$ is modest, and/or signal strength is low. Our results imply that selection methods are unlikely to recover the true model in such settings. However, because a "true model" does not really exist in applications [10,15], for descriptive purposes, a model that captures the data reasonably well may suffice.

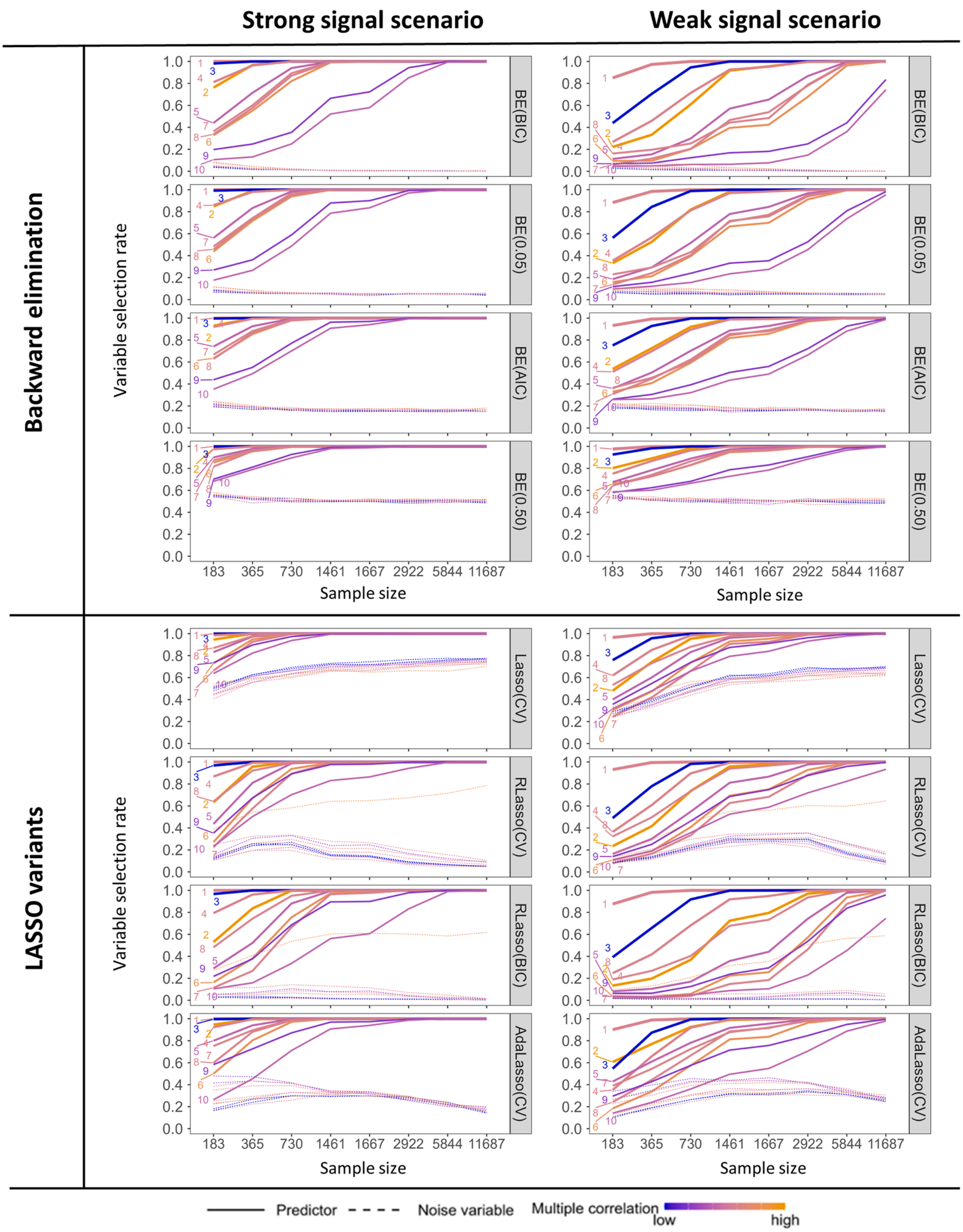


**Figure 1.** Logistic regression with event rate 0.3: For backward elimination and Lasso variants, variable selection rates (VSRs) of the 10 predictors (true positive rates) are shown with solid lines, and VSR of the 10 noise variables (false positive rates) are shown with dashed lines. Predictors are ranked 1—10 by their true standardized coefficients, with 1 indicating the strongest predictor. For each variable, line color indicates its multiple correlation with all other variables (the coefficient of determination from regressing that variable on all others), with the lowest value being 0 and the highest value being 0.56. Left panel: strong signal scenario $\left(R^2_{CS} = 0.40\right)$, right panel: weak signal scenario $\left(R^2_{CS} = 0.13\right)$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

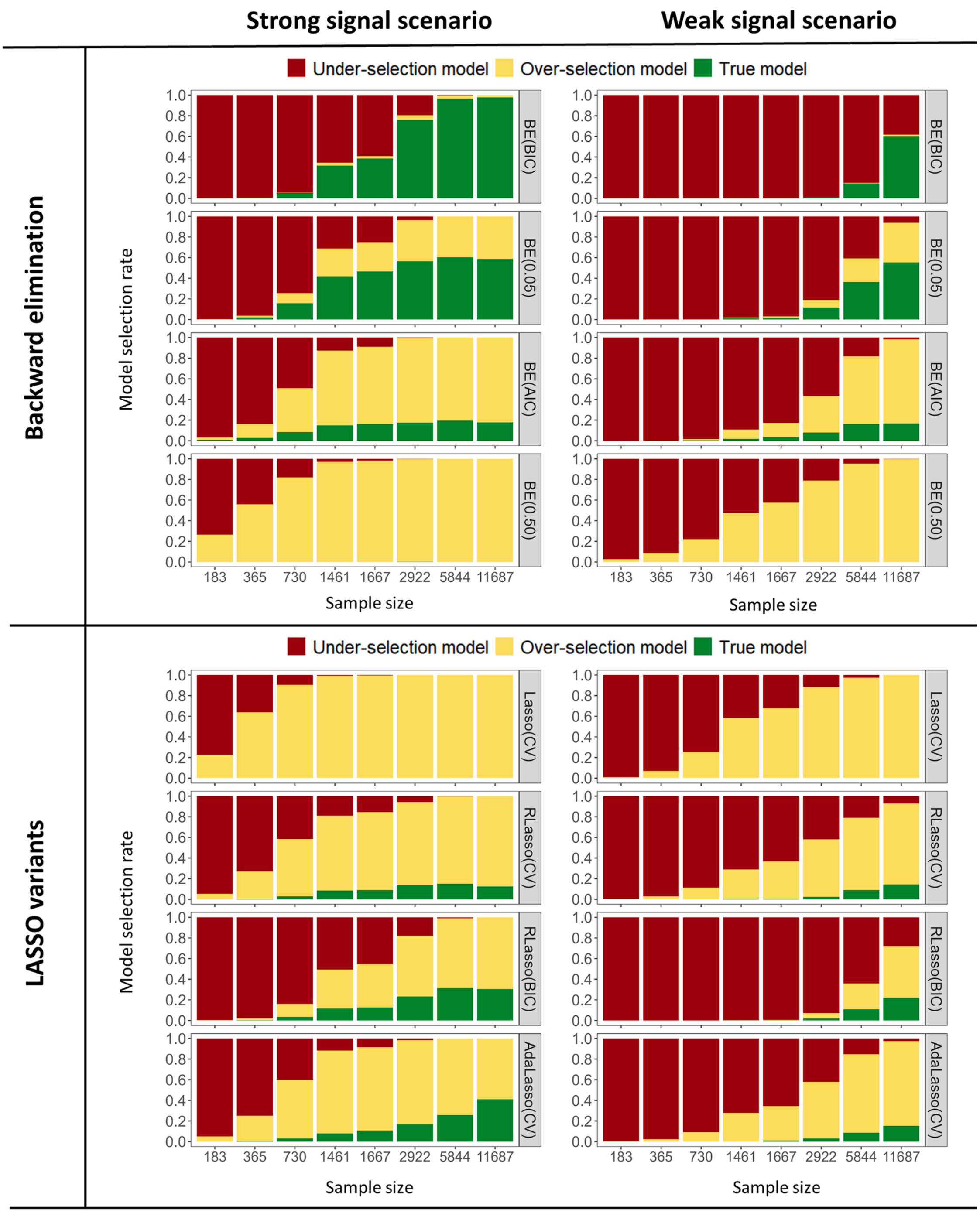


**Figure 2.** Logistic regression with event rate 0.3: model selection rate (MSR) of the true model, overselection models, which include all predictors plus at least one noise variable, and underselection models, which miss at least one predictor. Left panel: strong signal scenario $\left(R^2_{CS} = 0.40\right)$, right panel: weak signal scenario $\left(R^2_{CS} = 0.13\right)$.

### *3.3. Bias of coefficients*

We defined *conditional* bias as the expected bias of a variable's coefficient, given its selection, and *unconditional* bias as the expected bias of the variable's coefficient regardless of selection (nonselection corresponds to a coefficient of zero). For predictors, (un)conditional bias was defined such that positive values indicate overestimation and negative values underestimation. For noise variables, any nonzero estimate indicates overestimation.

*Logistic regression with event rate 0.3* (Fig 3, Fig 4): In the strong signal scenario, FU shows small (un)conditional bias away from zero for predictors at lower $n$ due to finite-sample bias in maximum likelihood estimates [28]. For noise variables, conditional bias increases with stricter selection because selected noise variables must have shown

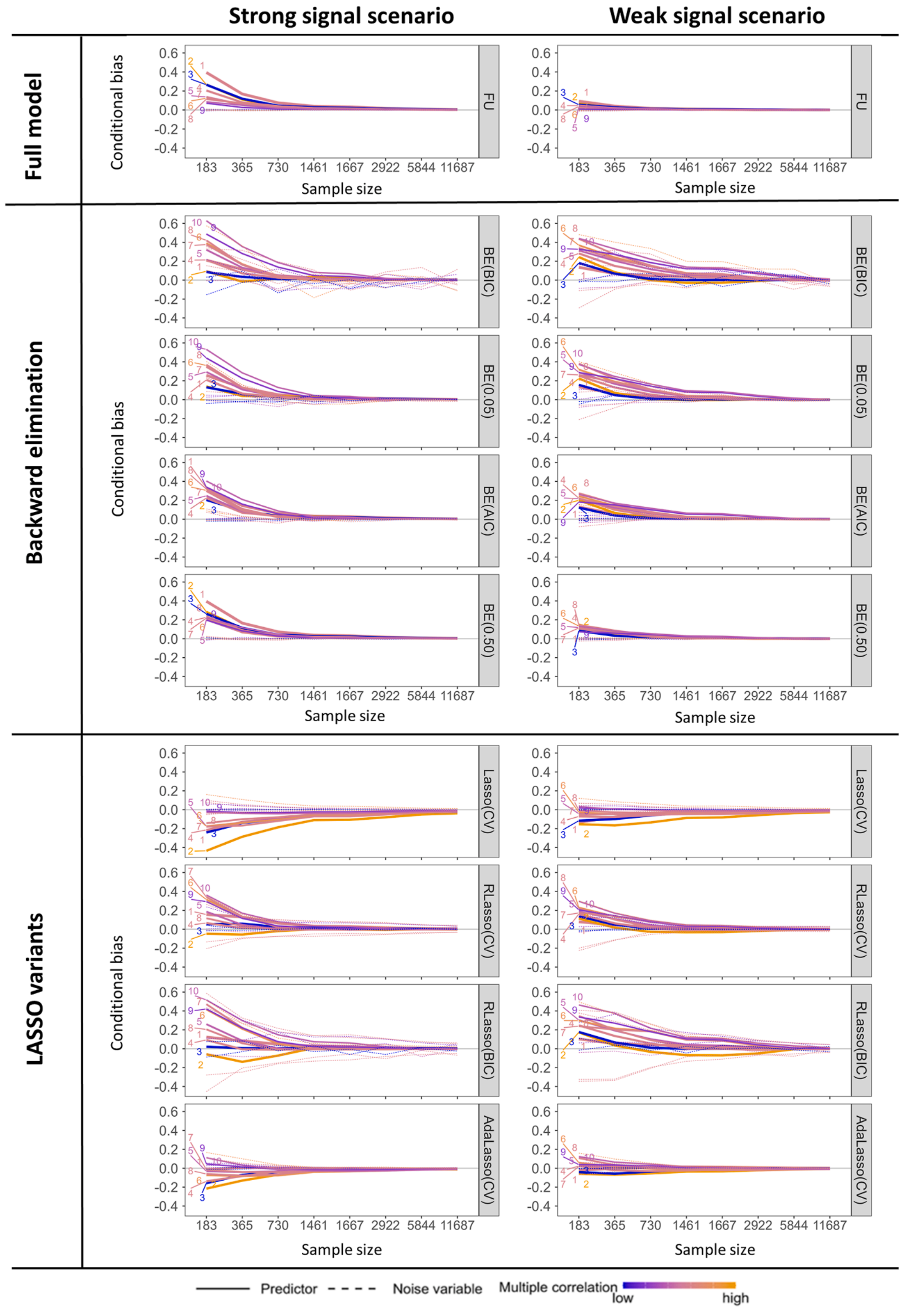


**Figure 3.** Logistic regression with event rate 0.3: conditional bias of regression coefficients. Predictors are shown with solid lines and noise variables are shown with dashed lines. For predictors, values >*rbin*0 indicate bias away from 0 (overestimation) and values <0 indicate bias toward 0 (underestimation). Predictors are ranked 1 to 10 by their true standardized coefficients, with 1 strongest. For each variable, line color indicates its multiple correlation with all other variables. Left panel: strong signal scenario $\left(R^2_{CS} = 0.40\right)$, right panel: weak signal scenario $\left(R^2_{CS} = 0.13\right)$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

stronger effects due to random variance in the respective simulated datasets in which they were selected. Conditional bias for predictors depends on the method. For BE, up to moderately large $n$, it is often away from zero, especially for weaker predictors and stricter selection, reflecting selection bias: rarely selected predictors must show inflated effects to be chosen. The Lasso induces conditional bias toward zero for predictors (shrinkage). The adaptive Lasso reduces shrinkage for strong predictors but slightly overestimates weaker ones (selection bias). The relaxed Lasso does not shrink coefficients, thus showing overestimation from selection bias. BIC tuning induces larger conditional bias than CV, reflecting stricter selection. As $n$ increases, conditional bias declines toward zero across methods, with slower convergence for BE(BIC), the Lasso, and the relaxed Lasso with BIC.

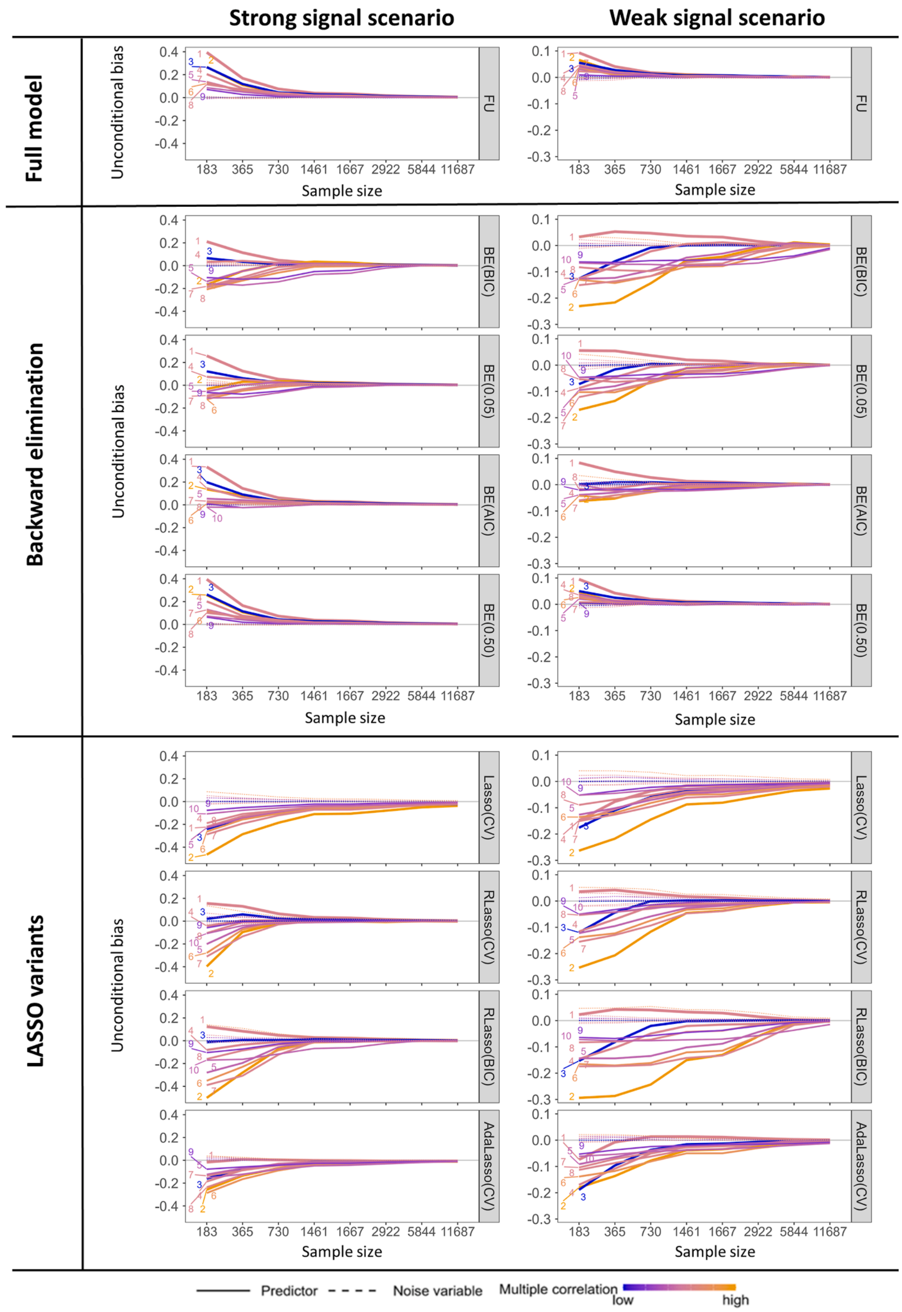


**Figure 4.** Logistic regression with event rate 0.3: unconditional bias of regression coefficients. Predictors are shown with solid lines and noise variables are shown with dashed lines. For predictors, values > *rbin*0 indicate bias away from zero (overestimation) and values < 0 indicate bias toward zero (underestimation). Predictors are ranked 1 to 10 by their true standardized coefficients, with 1 strongest. For each variable, line color indicates its multiple correlation with all other variables. Left panel: strong signal scenario $\left(R^2_{CS} = 0.40\right)$, right panel: weak signal scenario $\left(R^2_{CS} = 0.13\right)$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

In the weak signal scenario, coefficients are smaller by design to yield lower $R^2_{CS}$ [13]. Consequently, conditional (absolute) bias is lower, though *patterns* persist.

In both scenarios and across methods, unconditional bias is smaller than conditional bias for noise variables, since coefficients are set to zero when not selected. For predictors, when TPRs are below 1, unconditional bias shifts downward compared to conditional bias, moving from over- to under-estimation. Consequently, for BE with stricter criteria, unconditional bias becomes negative for several predictors. Unconditional bias for predictors in the weak signal scenario is not directly comparable to that in the strong signal scenario (different scales) but is more negative as fewer predictors are selected.

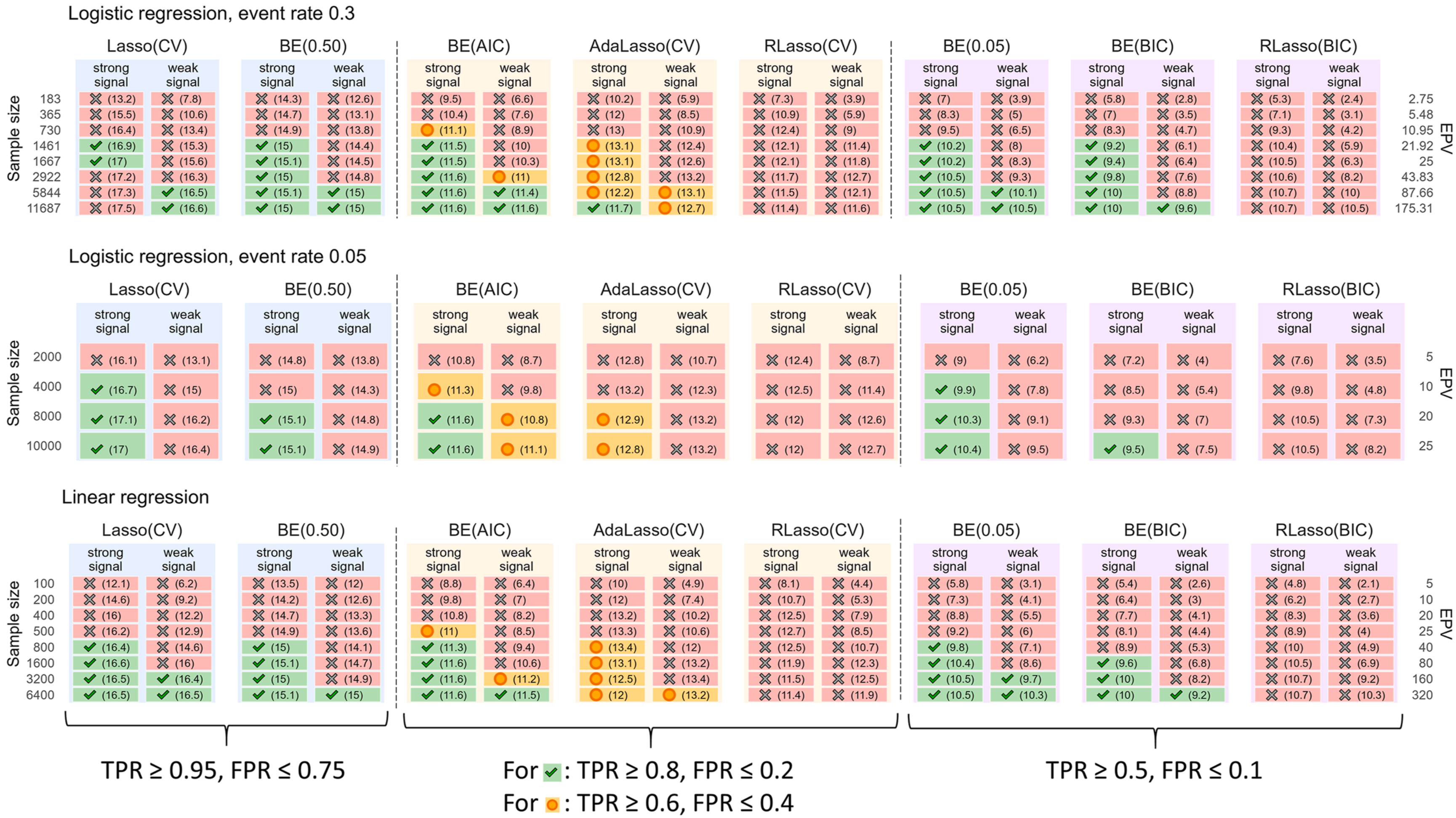


**Figure 5.** Decision framework: Variable selection methods are grouped into *mild selection* (left panel), *balanced selection* offering a compromise between TPR and FPR (middle panel), and *strict selection* prioritizing low FPR (right panel). Tiles indicate whether, for a given sample size/EPV and signal strength, each method meets the prespecified TPR/FPR criteria listed at the bottom of the figure, based on the simulation results. For balanced selection, both more stringent and more lenient criteria are shown. Green tiles with checkmarks indicate that the criteria are met, red tiles with crosses indicate that they are not met, and orange tiles with rings (middle panel) indicate fulfillment of the more lenient criteria. Next to each symbol, the mean number of selected variables (averaged over 2000 simulation repetitions) is shown in brackets. Strong signal strength refers to $R^2_{CS}$=0.40 (logistic, event rate 0.3), $R^2_{CS}$=0.16 (logistic, event rate 0.05), $R^2 = 0.45$ (linear); weak signal strength means $R^2_{CS} = 0.13$, 0.05, and $R^2 = 0.15$, respectively. To change the TPR/FPR criteria, we refer to Supplementary Material B. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

*Logistic regression with an event rate of 0.05* (Fig S7 and Fig S8): Conditional and unconditional bias mirror patterns of event rate 0.3 at comparable EPV.

*Linear regression* (Fig S9 and Fig S10): Results resemble logistic regression with event rate 0.3 at aligned *n*. As FU estimates are unbiased for all *n*, mild selection methods such as BE(0.50), which select many variables with moderate selection bias, exhibit low bias. Bias is directly comparable between strong and weak signal scenarios because true coefficients are identical. The higher bias under weak signals reflects the greater noise added to the linear predictor.

## 4. Discussion

Our findings have direct implications for epidemiological analyses, particularly those with modest EPV where data-driven variable selection is often applied post hoc. Rather than serving as a default modeling step, data-driven selection should be considered only when EPV and signal strength are sufficient. Figure 5 summarizes these considerations in a practical decision framework for applied epidemiologists and methodologists advising collaborative studies, clarifying when variable selection is defensible and when prespecification without selection is preferable. Based on our results and the existing literature, we recommend the following workflow.

1) Selection should begin with a carefully specified initial model, guided by prior knowledge and the literature [29], ideally before data collection. Although descriptive modeling does not seek a “true” model, it aims to capture underlying population signals. Our results show that data-driven selection rarely recovers the “true data-generating model”; researchers therefore cannot expect to discover the correct variables through selection alone. Even at high EPV, adequate performance requires that key variables are included initially, because no selection procedure can recover those excluded from the initial set.
2) Data-driven variable selection should be applied only under favorable conditions with respect to EPV and signal strength. To derive practical recommendations, we group methods into mild selection (BE(0.50), Lasso(CV)), strict selection (BE(BIC), BE(0.05), RLasso(BIC)), and selection balancing TPR and FPR (BE(AIC), RLasso(CV), AdaLasso(CV)). Method choice should reflect the desired TPR−FPR trade-off: mild methods prioritize retaining relevant variables, strict methods prioritize excluding noise variables, and balanced methods offer a compromise. Figure 5 shows, across sample sizes/EPV and signal strengths, whether methods meet predefined TPR/FPR criteria, defined as *all* predictors exceeding the TPR threshold and *all* noise variables remaining below the FPR threshold. The displayed thresholds provide a reasonable starting point.

   Overall, strong signals and/or sufficiently large EPV are required to meet these criteria, and coefficient bias is generally acceptable when they are fulfilled. For mild selection, the Lasso tends to select more variables and exhibits higher FPR than BE (0.50), failing the FPR criterion in some logistic regression settings. Among balanced methods, BE(AIC) performs best. For strict selection, BE(BIC) typically requires larger sample sizes than BE (0.05) to meet the TPR criterion. Relaxed Lasso variants fail mainly due to elevated FPR driven by highly correlated noise variables, but performance improves in simplified correlation structures (Fig S11). If different TPR/FPR criteria are required for a specific application, see Supplementary Material B for an interactive app to adjust these criteria (Fig 5).

   Our results align with previous recommendations of Heinze et al [10] but suggest that EPV thresholds should be higher, particularly under weak signals. As population-level signal strength is rarely known, estimation from prior studies may help. Otherwise, assuming a weak signal is a conservative and defensible choice.

3) When variable selection is used, it should be accompanied by stability analyses, especially at small-to-moderate EPV. Resampling approaches [10,30] evaluating VSR, MSR, and conditional coefficient bias can quantify selection-induced variability.
4) Transparent reporting is essential [6]. Studies should document the initial variable set and its rationale, EPV, the selection method used (including tuning or stopping criteria), and any assessment of stability.

We also assessed forward selection and forward selection with BE steps, which perform similarly to BE (Fig S12), but we do not recommend them as they can miss variables under certain correlations [18]. Augmented backward elimination [31], which combines BE with change-in-estimate, performs similarly to BE at larger sample sizes but shows higher TPR and FPR for smaller sample sizes for logistic regression (Fig S12). Supplementary Material C provides an app for more detailed exploration of all simulation results. Univariable selection is generally not recommended [32,33]. All findings and recommendations apply to descriptive modeling and should not be extrapolated to causal or predictive objectives, where different performance criteria and guidance are required [20,34,35]. Finally, although we used a realistic National Health and Nutrition Examination Survey-based correlation structure, robustness to alternative structures warrants further investigation, as does increasing the number of noise variables or considering alternative configurations of predictor effect sizes.

## 5. Conclusion

Selection performance depends critically on EPV and signal strength. For descriptive modeling, data-driven selection is defensible mainly in settings with high EPV and sufficiently strong signal, with the choice of method reflecting the desired TPR−FPR trade-off. Careful initial model specification, stability assessment, and transparent reporting remain essential regardless of method.

## CRediT authorship contribution statement

**Theresa Ullmann:** Writing − review & editing, Writing − original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Georg Heinze:** Writing − review & editing, Visualization, Supervision, Methodology, Conceptualization. **Michael Kammer:** Writing − review & editing, Visualization, Software. **Daniela Dunkler:** Writing − review & editing, Writing − original draft, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization.

## Ethics statement

No ethics approval was required, as this was a simulation study, which did not involve any human subjects or animals.

## Declaration of competing interest

The authors report no conflicts of interest related to this paper.

## Acknowledgments

We would like to thank the members of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (Topic Group 2 and the Publications Panel) for helpful comments. The STRATOS initiative is an international collaboration of biostatistical experts aiming to provide guidance in the design and analysis of observational studies. Topic Group 2 focuses on guidance for variable and functional form selection in multivariable analyses. In particular, we thank Willi Sauerbrei, Michal Abrahamowicz, Mariana Nold, and Frank Harrell. At the time of submission, Topic Group (TG) 2 of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative consisted of the following members (in alphabetical order): Michal Abrahamowicz, Harald Binder, Daniela Dunkler, Frank Harrell, Georg Heinze, Marc Henrion, Michael Kammer, Mariana Nold, Aris Perperoglou, Willi Sauerbrei, Matthias Schmid, and Theresa Ullmann. The group is co-chaired by Georg Heinze, Aris Perperoglou, and Willi Sauerbrei.

## Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112423.

## Data availability

All data and code are openly available on GitHub at https://github.com/thullmann/VariableSelection.

## References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodological) 1996;58(1):267−88.
[2] Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol 2009;24:733−6.
[3] Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 2010;8:20.
[4] Talbot D, Kubuta Massamba V. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol 2019;34(8):725−30.
[5] Vacheron CH, Friggeri A, Allaouchiche B, Maucort-Boulch D, Coz E. Quiet scandal: variable selection in three major intensive care medicine journals. Intensive Care Med 2021;47:1487−9.
[6] Pressat-Laffouilhère T, Jouffroy R, Leguillou A, Kerdelhue G, Benichou J, Gillibert A. Variable selection methods were poorly reported but rarely misused in major medical journals: literature review. J Clin Epidemiol 2021;139:12−9.
[7] Han MA, Kim HR, Yoon SE, Park SM, Kim B, Kim S-H, et al. How authors select covariates in the multivariate analysis of cancer studies in 10 oncology journals in Korea: a descriptive study. Sci Ed 2024;11(1):26−32.
[8] Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. In: Springer Series in Statistics. Cham: Springer International Publishing; 2015.
[9] Heinze G, Dunkler D. Five myths about variable selection. Transpl Int 2017;30(1):6−10.
[10] Heinze G, Wallisch C, Dunkler D. Variable selection−a review and recommendations for the practicing statistician. Biometrical J 2018;60(3):431−49.
[11] Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol 2018;98:133−43.
[12] Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res 2020;4:3.
[13] Ullmann T, Heinze G, Hafermann L, Schilhart-Wallisch C, Dunkler D, for TG2 of the STRATOS initiative. Evaluating variable selection methods for multivariable regression models: a simulation study protocol. PLoS One 2024;19(8):e0308543.
[14] Shmueli G. To explain or to predict? Stat Sci 2010;25:3.
[15] Carlin JB, Moreno-Betancur M. On the uses and abuses of regression models: a call for reform of statistical practice and teaching. Stat Med 2025;44:13−4.

[16] Shmueli G. To explain, to predict, or to describe: figuring out the study goal [commentary on ''On the Uses and Abuses of Regression Models'' by Carlin and Moreno-Betancur]. Stat Med 2025;44:13–4.

[17] Dyer BP. The distinction between causal, predictive, and descriptive research–there is still room for improvement. J Clin Epidemiol 2025; 189:111960.

[18] Nathan M. Why stepdown procedures in variable selection. Technometrics 1970;12(3):621–5.

[19] Meinshausen N. Relaxed lasso. Comput Stat Data Anal 2007;52(1): 374–93.

[20] Hastie T, Tibshirani R, Tibshirani R. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Stat Sci 2020;35(4):579–92.

[21] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101(476):1418–29.

[22] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38(11):2074–102.

[23] Sheppard JP, Stevens R, Gill P, Martin U, Godwin M, Hanley J, et al. Predicting out-of-office blood pressure in the Clinic (PROOF-BP): derivation and validation of a tool to improve the accuracy of blood pressure measurement in clinical practice. Hypertension 2016;67(5): 941–50.

[24] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey data. Hyattsville, MD: U.S. Department of Health, Human Services, Centers for Disease Control, and Prevention; 2026.

[25] Cario MC, Nelson BL. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Evanston, IL: Technical report, Department of Industrial Engineering and Management, Northwestern University; 1997.

[26] Ghosh S, Henderson SG. Behavior of the NORTA method for correlated random vector generation as the dimension increases. ACM Trans Model Computer Simulation (TOMACS) 2003;13(3):276–94.

[27] Kipruto E, Sauerbrei W. Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. PLoS One 2022; 17(10):e0271240.

[28] Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80(1):27–38.

[29] Hafermann L, Becher H, Herrmann C, Klein N, Heinze G, Rauch G. Statistical model building: background ''knowledge'' based on inappropriate preselection causes misspecification. BMC Med Res Methodol 2021;21:196.

[30] Wallisch C, Dunkler D, Rauch G, de Bin R, Heinze G. Selection of variables for multivariable models: opportunities and limitations in quantifying model stability by resampling. Stat Med 2021;40(2): 369–81.

[31] Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. PLoS One 2014;9(11):e113677.

[32] Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996;49(8):907–16.

[33] Ullmann T, Heinze G, Kappenberg F, Henrion M, Sauerbrei W, Leonhardt C-S, et al. The problem with univariable selection in regression modeling—and what to do instead. Vienna, Austria: Center for Medical Data Science, Medical University of Vienna; 2025.

[34] Luijken K, Groenwold RHH, van Smeden M, Strohmaier S, Heinze G. A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data. Biometrical J 2024;66(1):2100237.

[35] Kipruto E, Sauerbrei W. Evaluating prediction performance: a simulation study comparing penalized and classical variable selection methods in low-dimensional data. Appl Sci 2025;15(13):7443.