Received: 3 March 2023 | Revised: 31 January 2024 | Accepted: 7 February 2024

DOI: 10.1002/bimj.202300063

Biometrical Journal

RESEARCH ARTICLE

Interpretability of bi-level variable selection methods

Gregor Buch[1,2,3] | **Andreas Schulz**[1] | **Irene Schmidtmann**[2] | **Konstantin Strauch**[2] | **Philipp S. Wild**[1,3,4,5]

[1]Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[2]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[3]German Center for Cardiovascular Research (DZHK), Mainz, Germany

[4]Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[5]Institute of Molecular Biology (IMB), Mainz, Germany

Correspondence
Gregor Buch and Philipp S. Wild, Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany.
Email: Gregor.Buch@unimedizin-mainz.de and Philipp.Wild@unimedizin-mainz.de

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to confidentiality issues.

Abstract

Variable selection is usually performed to increase interpretability, as sparser models are easier to understand than full models. However, a focus on sparsity is not always suitable, for example, when features are related due to contextual similarities or high correlations. Here, it may be more appropriate to identify groups and their predictive members, a task that can be accomplished with bi-level selection procedures. To investigate whether such techniques lead to increased interpretability, group exponential LASSO (GEL), sparse group LASSO (SGL), composite minimax concave penalty (cMCP), and least absolute shrinkage, and selection operator (LASSO) as reference methods were used to select predictors in time-to-event, regression, and classification tasks in bootstrap samples from a cohort of 1001 patients. Different groupings based on prior knowledge, correlation structure, and random assignment were compared in terms of selection relevance, group consistency, and collinearity tolerance. The results show that bi-level selection methods are superior to LASSO in all criteria. The cMCP demonstrated superiority in selection relevance, while SGL was convincing in group consistency. An all-round capacity was achieved by GEL: the approach jointly selected correlated and content-related predictors while maintaining high selection relevance. This method seems recommendable when variables are grouped, and interpretation is of primary interest.

KEYWORDS

Bi-level selection, bootstrapping, group variable selection, interpretability

1 | INTRODUCTION

Scientific research questions often aim to identify clinically relevant characteristics that predict a response. When many predictors are potentially important, regularized regression is appealing as it penalizes all coefficients toward zero. This produces a parsimonious model that is easier to interpret than a model with all variables, as it separates relevant from irrelevant signals. This gain in interpretability is a common justification for using a selection method (Tibshirani, 1996), although parsimony is only one part of interpretability (Carvalho et al., 2019; Doshi-Velez & Kim, 2017). When predictors form groups based on their correlation or contextual similarity, related information is spread over numerous features. This makes selection at the group level more appropriate than identification of individual variables (Subrahmanya & Shin, 2010; Zaharieva et al., 2017). Example scenarios are when a set of features was collected with the same measurement instrument, was derived in different ways from the same data, or originated from the same experimental setting. Hence, the interpretability of a model increases when the following criteria are met:

(i) *Selection relevance*: Only features associated with the response are included because they are considered relevant and interpretability is enhanced when the generated model is parsimonious (Tibshirani, 1996; Hira & Gillies, 2015; Hesterberg et al., 2008).
(ii) *Collinearity tolerance*: Highly correlated predictors are treated alike since it is assumed that high correlation implies similarity in content, and interpretability is enhanced if those variables are selected jointly (Zou & Hastie, 2005; Bondell & Reich, 2008; Dormann et al., 2013).
(iii) *Group-level consistency*: Once a feature of a prespecified group is added to the model, all variables of that group are included, since it is likely that the predictors of a group are only meaningful together (Zaharieva et al., 2017; Yuan & Lin, 2006; Breheny & Huang, 2015; Gregorutti et al., 2015).

While the first criterion is a core competence of any selection technique, the second is possible with collinearity-tolerant approaches (Dormann et al., 2013), and the third defines group-level selection methods (Huang et al., 2012). Both concepts lead to more understandable results than a classical selection without such properties. In bi-level selection, a prespecified group formation is used in the selection process to identify both relevant variable groups and predictive variables within selected groups (Huang et al., 2009). Simpler techniques like the classical *least absolute shrinkage and selection operator* (LASSO) (Tibshirani, 1996) can mimic this to some extent if it is assumed that a group is selected as soon as at least one of its members is included in the model. It is still an unverified hypothesis that dedicated bi-level selection procedures are superior to such a strategy in terms of interpretability (Gregorutti et al., 2015; Lee et al., 2018; Afzal et al., 2021). Since they are not designed to optimize collinearity tolerance or group-level consistency, it remains unclear which characteristic should account for their supposed superiority.

To address this question, bootstrap samples of a real-world dataset with 1001 observations and 88 predictors were analyzed. The aim of the real-world application was to improve the understanding of the development and progression of autonomic dysfunction in patients with heart failure. The available feature space displayed interdependencies as variables were derived from the same devices or underlying measurements, leading to substantial correlations and contextual similarities among predictors. A composite of selection relevance, collinearity tolerance, and group-level consistency was used as a proxy criterion for interpretability. Using this criterion, the interpretability of *group exponential LASSO* (GEL) (Breheny, 2015), *sparse group LASSO* (SGL) (Simon et al., 2013), and *composite minimax concave penalty* (cMCP) (Breheny & Huang, 2009) was assessed and compared with that of classical LASSO. The effects of different group formation strategies (based on prior knowledge, correlation structure, or random assignment) and group sizes were examined in linear, logistic, and Cox regressions.

The first section outlines the dataset analyzed, the operationalization of interpretability, the origin of the group formations, and the methods under investigation. The subsequent section presents the results, followed by a discussion of the application scenarios in which the bi-level selection methods might be most appropriate.

2 | METHODS

2.1 | Dataset

All analyses were performed on a dataset of 1001 subjects with a set of heart rate variability parameters from recordings of a Holter ECG as part of the entry examination of the MyoVasc study (ClinicalTrials.gov Identifier: NCT04064450) (Göbel

TABLE 1 Measures of interpretability.

Interpretability dimension	Estimate	Scale	Summary measure
Selection relevance	BIC of a selected model predicting worsening of heart failure	0–100, rescaled according to the smallest and largest BIC of all generated models	Median
Collinearity tolerance	Mean absolute pair-wise correlation of selected variables	0–100, rescaled according to the lowest and highest mean absolute pairwise correlation present in the dataset	Median
Group consistency	Proportion of selected variables per selected group	Estimate on a scale between 0 and 100	Median

Abbreviation: Bayesian information criterion (BIC).

et al., 2021). This is a prospective cohort study investigating the development and progression of heart failure (Yancy et al., 2017; Yancy et al., 2013). These data were used to generate 1000 bootstrap samples in which variable selection methods were applied to identify relevant predictors for three dependent outcomes: (1) the biomarker NT-proBNP—the gold standard of laboratory biomarkers in heart failure—which is Gaussian distributed after natural logarithmic transformation, (2) the binary variable that distinguishes patients without heart failure from symptomatic heart failure, and (3) the time-to-event endpoint all-cause mortality. The methods were applied to the same set of 88 variables for each selection task, as listed in Tables S1 and S2 in the Supplemental Appendix. The analyses were repeated in a randomly drawn subsample of only 44 observations from the original analysis set. This allowed an investigation of the extent to which the results can be generalized to a high-dimensional setting with a low sample size. To investigate the calibration of selection approaches in the situation without any true association, analyses were repeated with permuted response variables. A simulation study was conducted to validate the results. For this purpose, the signal-to-noise ratio, the number of groups and variables, and the correlation structure were based on data from the MyoVasc study. Details on the data generation process can be found in the Supplemental Appendix (Simulation study S1). All calculations were performed using R Statistical Software v4.0.3 (R Core Team, 2021).

2.2 | Interpretability evaluation

Since interpretability is subjective, domain-specific, and not mathematically defined, it was necessary to rely on proxy criteria for an evaluation (Carvalho et al., 2019). To do so, we made use of the three criteria outlined in the introduction: the set of features to be interpreted should be reduced to a subset of predictive variables (I), with collinear variables being jointly selected (II), and predictors from the same group treated similarly (III).

The first criterion is referred to as selection relevance, the second as collinearity tolerance, and the last as group consistency. These criteria were considered different dimensions of interpretability and derived as stated in Table 1. For the first dimension, the models generated by the selection methods were used to predict the primary endpoint of the study, "worsening of heart failure" (Göbel et al., 2021). The goodness of these fits was determined using the Bayesian information criterion (BIC) (Schwarz, 1978), and the value was treated as selection relevance. The BIC was chosen to reward the generation of sparse predictive models. Based on the average absolute pairwise correlation of all selected features, collinearity tolerance was defined. If many highly correlated variables were selected, the average correlation of selected predictors increases and is reflected in this definition. Group consistency was determined by the proportion of selected variables per selected group. Thus, for example, a value of 50 means that half of all variables of the selected groups have been included in the model.

To make the dimensions comparable, they were scaled in a range from 0 to 100, where the larger the value, the higher the interpretability. Selection relevance was scaled using the largest and smallest observed BIC of all generated models, while for collinearity tolerance, the highest and lowest observed mean absolute pairwise correlations in the dataset were defined as maxima and minima. Group consistency is already on the correct scale.

The performance of the methods in the three dimensions of interpretability was calculated in all bootstrap samples, and the median of this distribution was used as a summary measure.

To reduce the dimensions to one value, the summary measures were combined with a weighted mean. The weights were chosen according to the summary measures so that the resulting score was not dominated by one dimension, as shown in Equation (1).

$$w_i = \frac{v_i^{-1}}{v_1^{-1} + v_2^{-1} + v_3^{-1}}, \tag{1}$$

where w_i denotes the weight of the ith interpretability dimension and $\boldsymbol{v}$ is a 1×3 vector containing the summary measures of the three dimensions.

For the comparison, a variable group was considered selected if at least one of its members was included in the model. The models formed were considered equivalent if the same predictors had a nonzero coefficient, whereas an empty model meant that no variable was selected.

2.3 | Group formation concepts

In the analyses, group formations were used in the selection processes based on domain knowledge (i.e., knowledge driven) or correlation structure (i.e., data driven) to group the predictors and evaluate the effects of different groupings on the selection results. Random group assignment was also performed to examine the effects of misclassified or uninformative grouping.

Regardless of the particular analysis, age, and sex were categorized as demographic characteristics, a variable group that did not change across groupings, so only the remaining 86 were grouped differently.

2.3.1 | The knowledge-driven group formation

The 86 variables could be divided into five nonoverlapping groups based on their similarities in content, as indicated in the Supplemental Appendix (Table S1): Eight features were combined into one variable group because they contained information on blood pressure. Another 16 predictors could be linked because they were derived from the same stress test. The remaining variables aggregate the intervals between successive heartbeats (R-R intervals) in three different domains, so that they were stratified into the following groups: Time domain (12 variables), frequency domain (31 variables), and nonlinear domain (19 variables) (Rodriguez-Linares et al., 2020).

2.3.2 | The data-driven group formation

To achieve a data-driven formation of variable groups, the features were grouped by Ward's hierarchical clustering method based on their Euclidean distance, using the implementation of the R package stats. This required determining the number of clusters, resulting in different group sizes for a fixed number of predictors. The 86 features were organized into data-driven formations, with the number of clusters corresponding to the Fibonacci sequence from 3 to 55 (Supplemental Appendix, Table S2). For the comparison between knowledge-driven, data-driven, and random assignments, the grouping with five clusters is of particular importance since these classifications have the same number of groups. The data-driven formation of five clusters is, therefore, presented in Tables S1 and S2 of the Supplementary Appendix.

2.3.3 | The random group assignment

To better assess the relevance of group formation, the assignment of knowledge-driven formation was permuted and the resulting structure was used as another group formation (Supplemental Appendix, Table S1). This classification helped to investigate whether data- or knowledge-driven grouping leads to different results than random and hence noninformative, group formation.

2.4 | Selection methods

All selection techniques discussed in this work use a penalty term added to the loss function of each task as given in Equation (2).

$$\mathrm{Q}\ (\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) = \mathrm{L}\,(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) + \mathrm{P}\,(\boldsymbol{\beta},\lambda)\,, \tag{2}$$

where $\boldsymbol{X}$ is a $n \times p$ matrix of J variable groups $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_j$ and K_j denotes the size of group j. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J)'$ with $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jK_j})'$ be the coefficient vector of group j.

A previously conducted systematic literature review from our group identified three frequentist penalty-based bi-level selection methods with implementation in R for Gaussian, binomial, and time-to-event response types (Friedman et al., 2010): SGL, GEL, and cMCP, implemented in the SGL (Simon et al., 2018) and grpreg (Breheny, 2015; Breheny & Huang, 2009) packages. These approaches were compared with the classical LASSO from the glmnet (Friedman et al., 2010; Simon et al., 2011) package, which is often treated in the literature as the gold standard for variable selection (Bhadra et al., 2019; Collignon et al., 2018; Soret et al., 2018) and is defined as follows:

$$\mathrm{P}^{\mathrm{LASSO}}\,(\boldsymbol{\beta}|\lambda) = \lambda\|\boldsymbol{\beta}\|_{1.} \tag{3}$$

2.4.1 | Group exponential LASSO

The GEL method was introduced in 2015, making it the newest bi-level selection method in this investigation. The method multiplicatively combines an exponential penalty at the group level with the LASSO penalty at the variable level and is defined as follows:

$$\mathrm{P}^{\mathrm{GEL}}(\boldsymbol{\beta}|\lambda,\tau) = \sum_{j=1}^{J} \frac{\lambda^2}{\tau}\left(1 - \exp\left(-\frac{\tau\|\boldsymbol{\beta}_j\|_1}{\lambda}\right)\right) \tag{4}$$

Due to the exponentiation of the group signal, the group level is given special weight in the selection process depending on the tuning parameter τ. This parameter controls the influence of the group information on the selection of each variable and can take values between 0 and 1. For a value of τ close to 0, the influence of the group signal is dampened, resulting in LASSO-like selections, while for values near 1, the emphasis in the selection process is on the group signals.

The recommended default value for τ of 1/3 was used for the analyses (Breheny, 2015).

2.4.2 | Sparse group LASSO

The most commonly cited bi-level selection method in the comparison is SGL, whose penalty corresponds to an additive combination of LASSO and group LASSO (Yuan & Lin, 2006) according to Equation (5).

$$\mathrm{P}^{\mathrm{SGL}}\ (\boldsymbol{\beta}|\lambda,\alpha) = \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\sum_{j\,=\,1}^{J}\|\boldsymbol{\beta}_j\|_2\right). \tag{5}$$

The tuning parameter α is a mixing parameter, where $(1-\alpha) * 100$ can be interpreted as the percentage relevance of the group information in the selection process. An α close to 1 yields LASSO-like results, where the collective signal of a group plays little role in the selection of a single variable, while a value near 0 yields group LASSO-like results since the group signal is exploited more.

In the R package SGL, a default of 0.95 is set for α, which leads to a very restrained use of the group structure. To evaluate the performance of the method, α was set to 0.5 in our analyses, which provides the maximum contrast to LASSO and group LASSO.

2.4.3 | Composite minimax concave penalty

Of the methods to be compared, cMCP is the only one that can estimate the coefficients unbiased despite their shrinkage. This is achieved by the *minimax concave penalty* (MCP) (Zhang, 2010) that reduces the penalty on the coefficients of predictors according to their magnitude in the regularization process.

The MCP is multiplicatively combined in cMCP and is, therefore, applied at both the intergroup level and the intragroup level.

$$\mathrm{P}^{\mathrm{cMCP}}\left(\boldsymbol{\beta}|\lambda,\ \gamma_1,\gamma_2\right)=\sum_{j=1}^{J}P^{MCP}_{\lambda,\gamma_1}\left(\sum_{k=1}^{K_j}P^{MCP}_{\lambda,\gamma_2}\left(\left|\beta_{jk}\right|\right)\right). \tag{6}$$

$$\mathrm{P}^{\mathrm{MCP}}\left(\theta|\lambda,\gamma\right)=\begin{cases}\lambda\left|\theta\right|-\frac{\theta^2}{2\gamma} & \text{if } \left|\theta\right|\le\gamma\lambda,\\ \frac{1}{2}\gamma\lambda^2 & \text{if } \left|\theta\right|>\gamma\lambda\end{cases}. \tag{7}$$

For the analyses, the tuning parameters γ_1 and γ_2, which control for the strength of the penalty reduction, were set to 3 as recommended by Breheny and Huang (2009).

Since group size is accounted for in the penalty term in Equation (6), the effect that a group's collective signal has on the selection of its members changes as a function of group sizes: the larger a group is, the weaker the effect, and vice versa.

2.4.4 | Values of the tuning parameters

For all approaches, the tuning parameter λ of the respective penalty was set to the largest value within one standard error of the minimum 10-fold cross-validated error term. Other parameters were set to default or recommended values, as described in the previous section.

To evaluate the effects of the tuning parameters of the bi-level selection methods on the results, subsequent analyses were performed with varying values for τ, α, and γ. For τ, the sequence 1/3, 1/30, and 1/300 was chosen, a trend that gradually leads to LASSO-like results from GEL. For α, values of 0.25, 0.5, and 0.75 were defined, to gradually produce more results like LASSO from SGL. The values for γ were set to 3, 30, and 300, a sequence that progressively limits the influence of the parameter and likewise leads to increasingly LASSO-like results from cMCP. These analyses were conducted using the data-driven group formation with 14 groups. This formation was chosen because it corresponds to the center of the Fibonacci sequence used to create the data-driven groupings.

3 | RESULTS

In the original analysis sample, the full model with all predictors achieved an R^2 of 0.45 in linear regression, an area under the receiver operating characteristic curve (AUC) of 0.85 in logistic regression, and a C-index of 0.86 in Cox regression. On average, the predictors correlated between 0.077 and 0.56 within knowledge-based groups, between 0.047 and 0.84 within data-driven groups, and between 0.11 and 0.14 for the random grouping. The mean pair-wise correlation of all predictors was 0.14.

Considerable differences between the selection approaches were apparent in their sparsity at the variable and group levels: regardless of group formation concepts (Supplemental Appendix, Table S3), the number of variable groups (Supplemental Appendix, Table S5), and values for the tuning parameters (Supplemental Appendix, Table S7), GEL resulted in the sparsest selection at the group level. This is evident in the case of four variable groups, where GEL selected one group on average, while other methods generally selected all groups. In contrast, cMCP often built the most parsimonious models at the feature level. Similar, but slightly less parsimonious was LASSO, while SGL invariably formed the largest

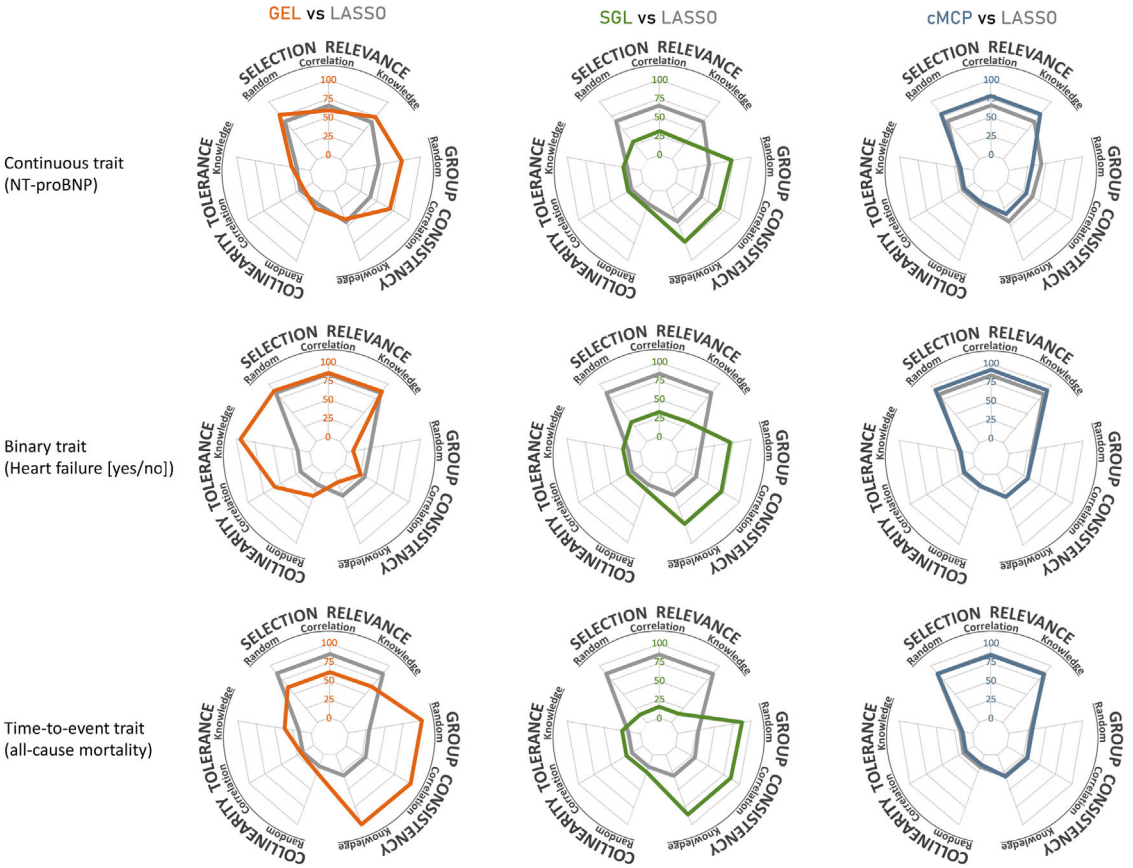

FIGURE 1 Effect of group formation concept on interpretability dimensions. Spider plots visualizing the effect of group formation concepts (based on knowledge, correlation, or randomization) on interpretability dimensions. The performance of the selection methods is described by the median values derived from 1000 bootstraps. Selection relevance is scaled from 0 to 100, according to the largest and smallest BIC of all generated models. Collinearity tolerance is scaled from 0 to 100, according to the lowest and highest absolute pairwise correlation present in the dataset. The center of the spider plot represents low interpretability, while the edges represent high interpretability. Applied selection methods: group exponential LASSO (GEL), sparse group LASSO (SGL), composite minimax concave penalty (cMCP), least absolute shrinkage, and selection operator (LASSO) as reference.

models. Overall, the foci set by the methods can be considered method-specific, since these remained fairly constant across the various applications.

Empty models were also built in the situations studied, especially in the time-to-event analyses. This occurred most frequently with GEL, which produced 251 empty models from 1000 bootstraps with data-driven group formation of 22 groups. For this selection procedure, GEL generated only 15 different models across all bootstraps, proving to be more consistent than SGL and LASSO, which formed a separate model in almost every bootstrap.

When grouping was held constant but method-specific tuning parameters were varied, sparsity at the group and variable levels changed most for GEL, followed by SGL, while little shifts were discernible for cMCP.

3.1 | Effect of group formation concepts on interpretability

Since several strategies for grouping predictors are possible, the influence of formation concepts on the defined dimensions of interpretability was investigated. For this comparison, we use the knowledge-driven group formation from Section 2.3.1, the data-driven group formation with five clusters from Section 2.3.2, and the random group assignment from Section 2.3.3. The results are visualized in Figure 1 for the three dependent variables with spider plots. The underlying values can be found in Table S4 of the Supplemental Appendix. For an alternative visualization that plots the performance of all methods for a trait, see Supplemental Appendix, Figure S1. As expected, the performance of LASSO did not change when the formation concept was modified, since this method does not account for the group information in the selection

TABLE 2 Interpretability of selection methods by group formation concept.

Dependent variable	Group formation concept	Selection method			
		GEL	SGL	cMCP	LASSO (ref.)
Continuous (NT-proBNP)	Knowledge-driven	**38.56**	37.71	32.96	34.51
	Data-driven	**37.8**	36.24	31.72	33.21
	Random	**48.28**	37.53	32.35	34.2
Binary (Heart failure [yes/no])	Knowledge-driven	**72.44**	38.27	34.77	34.25
	Data-driven	**55.17**	37.64	34.87	34.18
	Random	37.42	**37.91**	35.59	34.07
Time-to-event (all-cause mortality)	Knowledge-driven	**57.8**	38.93	32.59	33.99
	Data-driven	**48.22**	38.46	32.34	34.10
	Random	**51.4**	38.9	32.89	33.38

Note: Interpretability of the methods is described by the weighted mean of the performance in selection relevance, group consistency, and collinearity tolerance (see Table 1). The best results per group formation concept are shown in bold.Abbreviations: cMCP, composite minimax concave penalty; GEL, group exponential LASSO; LASSO, ref: least absolute shrinkage and selection operator as reference; SGL, sparse group LASSO

process. However, it is also evident from the plot that the effect of varying the grouping while keeping the number of groups constant barely affected the performance of SGL and cMCP.

Only GEL was influenced by the adjustments to any discernible degree. Its performance varied considerably across the three responses and across groupings in the interpretability dimensions. While GEL's selection relevance always remained at a similar level to LASSO and changed little across group formation concepts, the profile of GEL exhibited a focus on group consistency or collinearity tolerance depending on the response variable. Performance in these two dimensions sometimes changed sharply with variations in group formation concept.

With respect to group consistency, GEL achieved a performance of 100 in the time-to-event analysis, indicating that GEL selected groups completely. Since no other method reached such values, this appears to be a method-specific property. GEL was also the only approach that achieved an excellent value in collinearity tolerance, even if only in one situation. In particular, GEL was convincing in the analysis of the binary-dependent response with the highest value for collinearity tolerance of 95 (Supplemental Appendix, Table S4). This result was obtained because GEL frequently selected two heart rate reserve variables together in this scenario, which are correlated at 0.95 in the dataset.

The profiles of the other methods had a more pronounced focus than that of GEL. SGL demonstrated an emphasis on group consistency that was accompanied by a reduction in selection relevance. In particular, survival analysis showed near-perfect group consistency, but compared with the parsimonious selection of GEL, SGL achieved this by generating results resembling a full model (Supplemental Appendix, Table S3, and Figures S2–S4). The selection of almost all variables led to high group consistency but was represented by a low BIC, which explains the poor performance in selection relevance.

The performance of cMCP compared well with that of LASSO across all scenarios: both techniques performed worse than other methods in group consistency and collinearity tolerance (Supplemental Appendix, Table S4), but often convinced in selection relevance.

The reduction of the values of the interpretability dimensions to a single one is shown in Table 2. Across scenarios, GEL, followed by SGL emerged as the most interpretable method, while cMCP barely differed from LASSO. In some situations, SGL performed slightly better or equal to GEL but never came close to GEL's peak score of 72, which was achieved in the classification task with knowledge-driven group formation. As can be appreciated in the spider plots, the performance of GEL showed the highest variability, which is reflected in the variability of the interpretability score. All other methods showed much lower variability in their performance.

Selection with knowledge-based groupings often produced the best results and was superior to random assignment more often than data-driven formatting.

3.2 | Effect of group size on interpretability

The effect of a change in group size on dimensions of interpretability is shown in Figure 2 for the three target variables. Here, the various data-driven groupings are described in Section 2.3.2. were applied. In each panel of the diagram, the

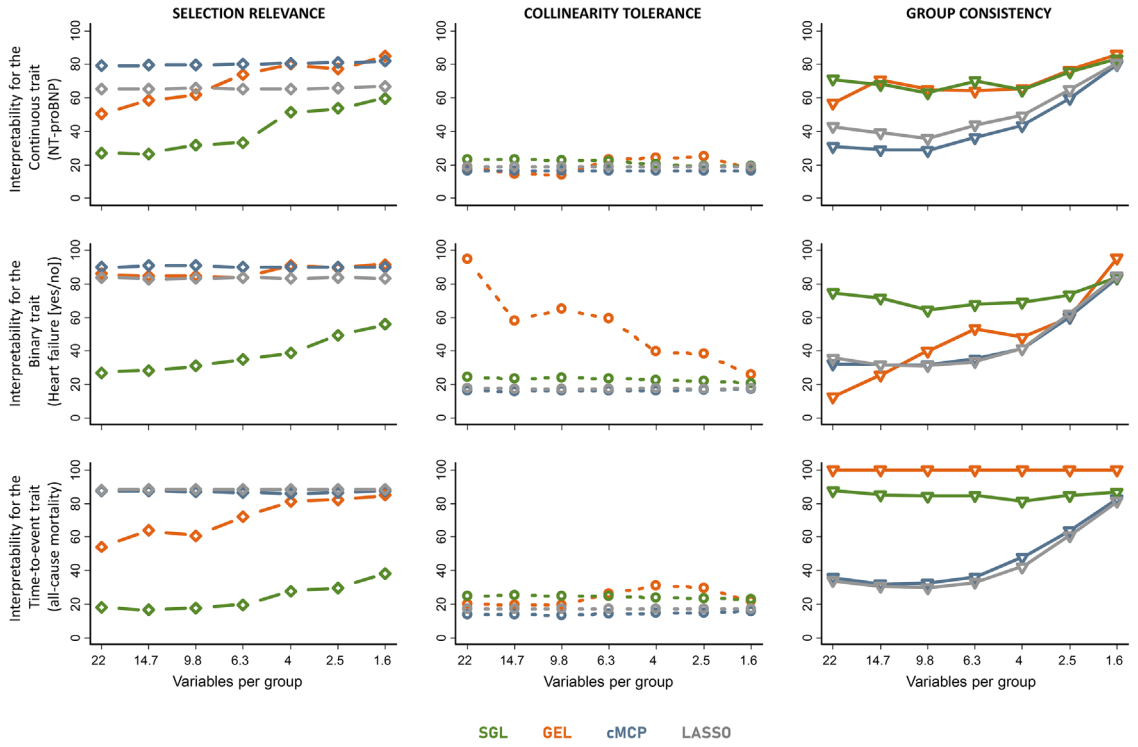

FIGURE 2 Effect of the number of variables per group on interpretability dimensions. Connected scatterplots visualizing the effects of variable group sizes on interpretability dimensions. The performance of the selection methods is described by the median values derived from 1000 bootstraps. Selection relevance is scaled from 0 to 100, according to the largest and smallest BIC of all generated models. Collinearity tolerance is scaled from 0 to 100, according to the lowest and highest absolute pairwise correlation present in the dataset. High values indicate high interpretability. In each situation, there are 88 variables that are distributed into different numbers of groups based on their correlation (data-driven group formation). Applied selection methods: group exponential LASSO (GEL), sparse group LASSO (SGL), composite minimax concave penalty (cMCP), least absolute shrinkage, and selection operator (LASSO) as reference.

performance of the methods by a variation of group size is presented. The values on which the plot is based are provided in the Supplemental Appendix (Table S6).

As with the spider plots, the performance of LASSO did not change for selection relevance and collinearity tolerance because this method does not include group information in the selection process. For group consistency, there was an increasing trend for smaller groups. This is due to the fact that in such designs the selection of a single predictor may already imply the selection of a large part of a group. Accordingly, the trend from LASSO provides an indication of how group consistency shifts based on group size alone.

Comparable to the results from the last section, limited variation in cMCP models was again found with a change in group size compared to LASSO. Consequently, cMCP showed convincing results on selection relevance but performed inferior in the other dimensions. The opposite was true for SGL, which performed worst on selection relevance but almost invariably performed best or second best on the other two dimensions. As in the previous analyses, SGL was strong on group consistency, reflecting its tendency to build large models (Supplemental Appendix, Table S5 and Figures S6–S8). The only method that could outperform SGL in this dimension was GEL. Yet, GEL achieved this performance with better selection relevance and partial collinearity tolerance than SGL. In addition, GEL showed unique performance in the classification task with 56 groups, where it scored highest in all dimensions.

Despite this superiority, GEL models exhibited the greatest variability: GEL was the only method whose trajectories did not always show a clear trend and the only one that performs better or worse than LASSO depending on group size, as in the case of group consistency for the binary dependent variable.

However, as interpretability in Table 3 shows, GEL mostly scored highest on the composite criteria. In one exception, GEL was outperformed by SGL, although the difference was marginal. Compared with LASSO, SGL appeared to perform better in situations with few large groups, whereas the performance of cMCP increased the smaller the groups. However, both approaches only marginally outperformed LASSO in terms of interpretability, while GEL was superior to LASSO regardless of the average number of features per group.

TABLE 3 Interpretability of selection methods by the number of variable groups.

		Selection method			
Dependent variable	**Data-driven group formation**	**GEL**	**SGL**	**cMCP**	**LASSO (ref.)**
Continuous (NT-proBNP)	Avg. no. variables per group: 22	35.37	**37.33**	32.51	34.43
	Avg. no. variables per group: 14.67	**38.74**	36.34	32.04	33.45
	Avg. no. variables per group: 9.78	**37.30**	35.66	31.99	32.63
	Avg. no. variables per group: 6.29	**44.27**	37.81	34.17	34.74
	Avg. no. variables per group: 4	**46.40**	38.51	36.33	36.23
	Avg. no. variables per group: 2.51	**49.33**	41.52	41.01	40.62
	Avg. no. variables per group: 1.57	**49.52**	44.90	46.80	45.24
Binary (Heart failure [yes/no])	Avg. no. variables per group: 22	**70.32**	38.97	34.85	35.42
	Avg. no. variables per group: 14.67	**54.08**	37.80	34.80	34.00
	Avg. no. variables per group: 9.78	**61.85**	36.61	34.94	33.92
	Avg. no. variables per group: 6.29	**62.24**	38.08	35.70	34.64
	Avg. no. variables per group: 4	**51.86**	38.66	37.50	36.87
	Avg. no. variables per group: 2.51	**54.15**	41.62	42.89	42.61
	Avg. no. variables per group: 1.57	**57.78**	45.14	49.61	49.02
Time-to-event (all-cause mortality)	Avg. no. variables per group: 22	**48.81**	41.02	33.91	35.34
	Avg. no. variables per group: 14.67	**50.37**	40.21	32.93	34.59
	Avg. no. variables per group: 9.78	**49.90**	40.09	32.81	34.34
	Avg. no. variables per group: 6.29	**55.52**	40.47	34.06	35.20
	Avg. no. variables per group: 4	**59.92**	40.53	37.53	37.72
	Avg. no. variables per group: 2.51	**59.35**	41.73	42.03	42.93
	Avg. no. variables per group: 1.57	**55.70**	43.51	48.03	48.51

Note: Interpretability of the methods is described by the weighted mean of the performance in selection relevance, group consistency, and collinearity tolerance (see Table 1). The best results per average (Avg.) number of (no.) variables per group are shown in bold.Abbreviations: cMCP, composite minimax concave penalty; GEL, group exponential LASSO; LASSO, ref: least absolute shrinkage and selection operator as reference; SGL, sparse group LASSO.

3.3 | Effect of tuning parameters on interpretability

Since the values of the tuning parameters (τ, α, and γ) of the bi-level selection methods had an influence on the selection, they were varied in the further analyses. The data-driven group formation with 14 groups was used here, as explained in Section 2.4.4. The results are shown in Figure S2 and presented with the data table in the Supplemental Appendix Table S9.

Changing the initial setting affected the selection process of cMCP slightly, whereas significant differences were seen for SGL and GEL. Adjustments of α correspond to a rebalancing of the emphasis of SGL between group consistency and selection relevance: increasing α from 0.25 to 0.5 and 0.75 increased within-group sparsity (Supplemental Appendix, Table S8) and selection relevance, whereas collinearity tolerance remained at a constant level. Varying τ caused slight changes in selection relevance for GEL, while shifts in collinearity tolerance and group consistency were observed.

Table S9 in the Supplemental Appendix presents the interpretability of the approaches as a function of a change in the tuning parameters. GEL showed the best performance across selection tasks when τ was set to the default value, whereas SGL often demonstrated superiority when deviating from this default value. The variability of the results confirmed that small changes in the selection task setting can have a large impact on the performance of GEL, while the other methods produced relatively stable results.

3.4 | Sensitivity analyses

Results from the analyses with a reduced sample size ($N = 44$) can be found in the Supplemental Appendix in Table S10–S15. Due to the small sample size, the methods selected fewer variables and groups, and more often formed empty models. The ranking of the techniques in terms of their performance on the interpretability dimensions remained similar

to those analyses with the whole sample. However, the selection relevance of SGL improved in the analyses that considered different grouping concepts but still lagged behind all other methods. Reducing the group size again led to better group consistency in all approaches. Only GEL benefited less but was again able to achieve a group consistency of 100 for the time-to-event trait. In addition, performance on selection relevance of the techniques equalized as the groups became smaller. Thus, SGL almost reached the performance of the other methods. As with the full analysis set, GEL was the only method that achieved noteworthy high values on collinearity toleration. Depending on the trait, this was favored by larger or smaller groups. Varying the tuning parameters confirmed previous results, such as the high variability of GEL with a change in τ.

After permuting the dependent variable, the selection methods should no longer select any variables. This was verified in 1000 bootstrap samples, and the results can be found in Table S16 in the Supplemental Appendix. Across all selection tasks, LASSO generates empty models for the continuous trait in up to 93.5% of all bootstrap samples. Thus, LASSO clearly outperforms the bi-level selection methods, which achieved this in at most 19.8% of the cases. When utilizing knowledge-driven group formation for the binary trait, GEL generated the empty model in 68.3% of the repetitions. This was the best result obtained by bi-level selection methods. Unlike LASSO, the bi-level selection methods also tended to select more variables than in the nonpermuted situation.

The results of the simulation study to validate the findings can be found in Table S17 in the Supplemental Appendix. Familiar patterns could be validated in terms of sparsity. For example, SGL always selected the most variables and cMCP often selected the fewest variables, while GEL was characterized by clear sparsity at the group level. The results were also validated regarding performance in the interpretability dimensions. Again, cMCP was the best method for selection relevance, GEL was superior in group consistency and in collinearity tolerance, followed by SGL. The most notable differences were the performance of GEL for collinearity tolerance in linear regression and group consistency in logistic regression. In both cases, GEL was the worst method with real-world data, but the best in the simulation study. In logistic regression, GEL achieved the best possible group consistency, that is, all variables in a group were selected.

4 | DISCUSSION

By its very nature, variable selection serves to increase interpretability (Tibshirani, 1996). These methods reduce the set of characteristics to be examined so that a smaller, and therefore simpler, model can be understood. Depending on the setting, interpretability can be improved if collinearity between predictors is tolerated or variables belonging to a predefined group are selected together (Zou & Hastie, 2005; Yuan & Lin, 2006). In this work, three criteria were treated as dimensions of the interpretability of selection approaches and used for method evaluation. The aim was to compare implemented bi-level selection methods with classical LASSO in terms of interpretability. The results show that bi-level selection methods can improve interpretability beyond that of LASSO. In detail, however, assumptions about the grouping of predictors are necessary and method-specific properties have to be considered since not every bi-level selection method is appropriate for all applications.

In general, it is possible to use domain knowledge or the correlation structure of the data to form variable groups (He & Yu, 2010). Since the use of knowledge-driven model building usually resulted in the highest values for interpretability, this concept seems to be recommendable. This is largely in line with other findings in the literature (Buch et al., 2023). However, it must be taken into account that a random group assignment performed in some selection tasks better than the alternative groupings. This classification strategy was performed to assess the relevance of the formation concepts, and the results underlined their limitations: Considering groups in selection can be a hindrance if the formation is not based on information relevant to predict the response.

Compared to the effect of the group formation concept, the average number of predictors per group seems to influence the selection more strongly. While some methods showed little response to adjusting the group formation concept, a variation in the number of variables per group led to changed results for all methods. Depending on the group size, the methods investigated performed better or worse. This has relevant implications for application, as numerous groupings are possible for different datasets (Wu et al., 2019). In the case of -omics data (e.g. transcriptomics, proteomics), prior knowledge allows for the formation of large functional groups or the partitioning into smaller but more specific groups (Daneshgar et al., 2021; Denis et al., 2014; Forest et al., 2017; Qiu et al., 2015). The same is true when a data-driven grouping is conducted, for example, with hierarchical clustering, which allows the formation of different group sizes (Das et al., 2020; Mameli et al., 2021; Wang et al., 2021). Since the interpretability of most methods increased when smaller groups were built, such formations seem preferable to coarser groupings. Since the interpretability of LASSO also improved with

smaller groups, the results provide a rough outline of which bi-level selection method would be superior to LASSO for a given group size. The cMCP was only able to outperform LASSO in settings with smaller groups. Above an average group size of four features per group, the interpretability of cMCP in the linear regressions was lower than that of LASSO. Using cases with larger groups does not seem advisable. SGL performed better than LASSO when groups had an average group size greater than four. When groups were smaller, there was no advantage of SGL over LASSO.

Findings from previous comparisons align with our results (Belhechmi et al., 2020; Xiang et al., 2013; Liu et al., 2013; Matsui, 2015; Shen et al., 2021; Cai et al., 2019), but more targeted simulation studies could be useful to identify an optimal setting for GEL since no clear trend for this method emerged from our results. The distinct performance of SGL in group consistency can be confirmed with published simulation studies based on the model size formed (Belhechmi et al., 2020). The same is true for GEL pronounced sparsity at the group level as well as the lack thereof in cMCP, which has also been reported in other applications (Shen et al., 2021). The predictive performance of the methods in simulation studies supports the ranking of selection relevance postulated here, with cMCP at the top (Liu et al., 2013). Comprehensive simulation studies with knowledge- and correlation-based groups have already suggested that GEL tends toward collinearity-tolerant selection and the inability of the same for SGL and cMCP (Buch et al., 2023). Yet, even if interpretability could be improved, the integration of grouping is not generally advisable over a classical approach. Just because variables can be grouped, and this information can be considered in a selection process does not mean that this necessarily improves the selection. A meaningful application requires that the researcher has an interest in the group-level interpretation and has confidence in the meaningfulness and correctness of the grouping.

If this condition is met, GEL is recommended because it is superior to other bi-level selection methods and LASSO in terms of interpretability. This method can achieve high selection relevance but also exhibits high group consistency and collinearity tolerance: a unique all-round capability among the methods in comparison. Also unique is that GEL is very sensitive to changes in group formation, indicating that grouping has a strong influence on the selection process. Although this is a desirable feature of a bi-level selection method, the marked sensitivity of GEL could make replication of results or sensitivity analyses difficult. Moreover, the high-performance variability of GEL suggests that its superiority over other techniques may be less in situations other than those we studied.

The other methods mostly selected only a defined part of a variable group, whereas GEL showed a unique tendency to include variable groups completely in the model, especially in the survival analyses. This behavior is more reminiscent of a group-level selection method and thus represents a specific conception of "bi-level" selection. Depending on the situation, it may be desirable or inappropriate. It is beneficial when a group of variables is considered relevant if it contains many weak signals that together have explanatory power. However, it may also be inappropriate, as it could lead to the inclusion of variables in the model that have no association with the response but only belong to a group with predictive variables (Kwon et al., 2017). Because of this property, a high degree of confidence in the group information is required to use the approach meaningfully.

SGL is recommended as the second-best technique in comparison, but the method needs further development to improve the deficiencies in selection relevance. The approach has a clear focus on group consistency, which is rarely achieved to this extent by other methods. A clear disadvantage is its poor performance in selection relevance, which is mainly due to the tendency of SGL to produce very large models. Compared to classical LASSO, the question arises whether the gain in group consistency justifies the loss in selection relevance. This exchange can be optimized with the tuning parameter α in SGL. In our practical application, however, the adjustment of α did not lead to a satisfactory improvement in selection relevance. A possible starting point to refine the method would be to combine other penalty functions in the style of SGL instead of the group LASSO and the classical LASSO. A mixture of the group MCP (Wei & Zhu, 2012) and the classical MCP has already been presented (Liu et al., 2013), but a suitable implementation of this approach is lacking (Huang et al., 2012; Buch et al., 2023). The technique presented there seems to form more parsimonious models and responds to a change of α in a more recognizable way. By gradually increasing α, the influence of grouping can be varied such that variables in a group are gradually removed from the model. This could be considered as additional information for the interpretation of the results and makes the influence of grouping in the selection process more transparent.

Therefore, SGL seems to be particularly suitable for projects in which the group structure can be doubted and sensitivity analyses are needed to investigate the robustness of the results. Such sensitivity analyses could be performed with SGL with a variation of α.

The cMCP makes insufficient use of the group information provided to the approach and is therefore rarely preferable to LASSO in terms of interpretability. The method can be classified as the clear favorite for selection relevance but performs similarly to LASSO on the other dimensions of interpretability. The superiority of cMCP in selection relevance appears to be due to the MCP elements in the penalty term. It has already been shown that MCP has advantages over LASSO for

feature selection (Zhang, 2010). In parallel, group information is hardly used. Only in situations with small groups, the additional information seems to play a minor role (Breheny, 2015). A possible application of this method is to deescalate that signals from small groups are better selected at the group level, while for larger groups only the relevant elements should be selected.

We considered several dimensions of interpretability that are characteristic of implemented frequentist regularized regression techniques for bi-level selection. Alternative aspects of interpretability, which can be achieved by means of graphs, Shapley values, or important scores, have not been investigated. None of the implementations compared was able to provide valid confidence intervals for the coefficients—information that would otherwise also aid interpretation. Bayesian approaches could overcome this shortcoming (Cai et al., 2019; Xu & Ghosh, 2015). Alternatively, sample splitting (training and test dataset), resampling (i.e., bootstrapping) techniques (Meinshausen & Bühlmann, 2010), knockoffs (Barber & Candès, 2015), or penalty-specific (i.e., for LASSO or MCP) solutions could be used for postselection inference (Taylor & Tibshirani, 2018; Chai et al., 2019).

In addition to investigating the interpretability aspect, this paper was the first to analyze the influence of incorrect group information in the selection process of bi-level selection methods. Methods were identified that react sensitively (GEL) or robustly (cMCP) to erroneous groupings, and it was discussed how this can be handled using sensitivity analyses. To our knowledge, this is the first time that the influence of group size and definition (knowledge driven or data driven) on the performance of bi-level selection has been systematically analyzed and discussed. Another new aspect is the use of various group structures that are not just simulated but occur in reality to increase the relevance of these investigations for practical application. Previous applications in the literature have considered one group structure only, although real datasets can usually be divided into several alternative groups. Moreover, prior comparisons in this domain have omitted a holistic evaluation encompassing linear, logistic, and Cox regressions, while also neglecting the impact of tuning parameters. Consequently, this work serves as a vital complement to the existing comparisons and contributes essential aspects that reflect the demand for Phase III and IV studies in methodological research (Heinze et al., 2022). Notably, this research mitigates potential inventor bias and thus addresses the call for neutral comparison studies (Boulesteix et al., 2013).

Based on this method comparison, an attempt was made to provide recommendations for the application of the approaches studied in practice. Such statements are not applicable to all situations, and simplifications made are not appropriate for every research question. Instead, the postulated ranking is only a rough guideline. In particular, the aggregation of all interpretability dimensions to just one number can be designed in many ways. We used a weighted mean to constrain the otherwise dominant influence of selection relevance. Most methods were quite successful at selection relevance-that is, after all, their main competence. If the dimensions were combined unweighted, methods such as cMCP would appear better, even though they performed the worst in group consistency and collinearity toleration.

The results based on permuted response variables suggest that bi-level selection methods are more likely to select false positive features than classical LASSO. It is therefore advisable to use LASSO as a sensitivity analysis when bi-level selection methods are employed.

5 | CONCLUSIONS

Bi-level selection methods can significantly improve the interpretability of results compared to a single variable selection method like LASSO. In particular, the performance of GEL is convincing, as it selects correlated and substantively similar features together without showing substantial losses in selection relevance. The approach can, therefore, be recommended when predictors are grouped and increased interpretability is desired. An alternative could be SGL as the technique is versatile due to the function of its tuning parameter. The least suitable for improving interpretability seems to be cMCP since the method only appreciably considers information about small groups.

ACKNOWLEDGMENTS

This work is part of the dissertation of Gregor Buch and was supported by the DIASyM project, funded by the Federal Ministry of Education and Research (BMBF, grant No. 031L0217A) and the Center of Preventive Cardiology and Preventive Medicine of the University Medical Centre of the Johannes Gutenberg-University Mainz. Philipp. S. Wild is the principal investigator of the German Center for Cardiovascular Research (DZHK), principal investigator of the DIASyM research core (BMBF DIASyM research core (BMBF 031L0217A)), principal investigator of the Institute of Molecular Biology and was funded by the Ministry of Education and Research (BMBF 01EO1003 and 01EO1503).

The MyoVasc study was supported by funding from the German Center for Cardiovascular Research (DZHK), the Center for Translational Vascular Biology (CTVB) from the University Medical Center Mainz, and own funding. This analysis project received financial support from Bayer AG.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

All authors declare themselves to have nothing to disclose that could be perceived as a conflict of interest in the context of the present work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to confidentiality issues.

ORCID

Gregor Buch https://orcid.org/0000-0002-9963-1245

REFERENCES

Afzal, A. R., Yang, J., & Lu, X. (2021). Variable selection in partially linear additive hazards model with grouped covariates and a diverging number of parameters. *Computational Statistics*, *36*, 829–855. https://doi.org/10.1007/s00180-020-01062-3

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-AOS1337

Belhechmi, S., Bin, R. D., Rotolo, F., & Michiels, S. (2020). Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC Bioinformatics*, *21*, 1–20. https://doi.org/10.1186/s12859-020-03618-y

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, *34*, 405–427. https://doi.org/10.1214/19-STS700

Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, *64*, 115–123. https://doi.org/10.1111/j.1541-0420.2007.00843.x

Boulesteix, A.-L., Lauer, S., & Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PloS ONE*, *8*, e61562. https://doi.org/10.1371/journal.pone.0061562

Breheny, P. (2015). The group exponential LASSO for bi-level variable selection. *Biometrics*, *71*, 731–740. https://doi.org/10.1111/biom.12300

Breheny, P., & Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, *2*, 369. https://doi.org/10.4310/SII.2009.v2.n3.a10

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*, 173–187. https://doi.org/10.1007/s11222-013-9424-2

Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., & Wild, P. S. (2023). A systematic review and evaluation of statistical methods for group variable selection. *Statistics in Medicine*, *42*, 331–352. https://doi.org/10.1002/sim.9620

Cai, M., Dai, M., Ming, J., Peng, H., Liu, J., & Yang, C. (2019). BIVAS: A scalable Bayesian method for bi-level variable selection with applications. *Journal of Computational Graphical Statistics*, *29*, 1–38.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*, 832. https://doi.org/10.3390/electronics8080832

Chai, H., Zhang, Q., Huang, J., & Ma, S. (2019). Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Statistica Sinica*, *29*, 877.

Collignon, O., Han, J., An, H., Oh, S., & Lee, Y. (2018). Comparison of the modified unbounded penalty and the LASSO to select predictive genes of response to chemotherapy in breast cancer. *PloS ONE*, *13*, e0204897. https://doi.org/10.1371/journal.pone.0204897

Daneshgar, N., Baguley, A. W., Liang, P.-I., Wu, F., Chu, Y., Kinter, M. T., Benavides, G. A., Johnson, M. S., Darley-Usmar, V., Zhang, J., Chan, K.-S., & Dai, D.-F. (2021). Metabolic derangement in polycystic kidney disease mouse models is ameliorated by mitochondrial-targeted antioxidants. *Communications Biology*, *4*, 1–13. https://doi.org/10.1038/s42003-021-02730-w

Das, K., Kenny, A., & Solomon, D. (2020). Analyzing high-dimensional gene expression data by using the regularization. *In Joint Statistical Meetings*.

Denis, M., Enquobahrie, D. A., Tadesse, M. G., Gelaye, B., Sanchez, S. E., Salazar, M., Ananth, C. V., & Williams, M. A. (2014). Placental genome and maternal-placental genetic interactions: A genome-wide and candidate gene association study of placental abruption. *PloS ONE*, *9*, e116346. https://doi.org/10.1371/journal.pone.0116346

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Forest, M., Iturria-Medina, Y., Goldman, J. S., Kleinman, C. L., Lovato, A., Oros Klein, K., Evans, A., Ciampi, A., Labbe, A., & Greenwood, C. M. T. (2017). Gene networks show associations with seed region connectivity. *Human Brain Mapping*, *38*, 3126–3140. https://doi.org/10.1002/hbm.23579

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Göbel, S., Prochaska, J. H., Tröbs, S.-O., Panova-Noeva, M., Espinola–Klein, C., Michal, M., Lackner, K. J., Gori, T., Münzel, T., & Wild, P. S. (2021). Rationale, design and baseline characteristics of the MyoVasc study: A prospective cohort study investigating development and progression of heart failure. *European Journal of Preventive Cardiology*, *28*, 1009–1018. https://doi.org/10.1177/2047487320926438

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, *90*, 15–35.

He, Z., & Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, *34*, 215–225. https://doi.org/10.1016/j.compbiolchem.2010.07.002

Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., & Simulation Panel of the STRATOS initiative. (2022). Phases of methodological research in biostatistics—Building the evidence base for new methods. *Biometrical Journal*, *66*(1), 2200222. https://doi.org/10.1002/bimj.202200222

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and 1 penalized regression: A review. *Statistics Surveys*, *2*, 61–93. https://doi.org/10.1214/08-SS035

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*, 198363.

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *27*(4), 481–499. https://doi.org/10.1214/12-STS392

Huang, J., Ma, S., Xie, H., & Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, *96*, 339–355. https://doi.org/10.1093/biomet/asp020

Kwon, S., Ahn, J., Jang, W., Lee, S., & Kim, Y. (2017). A doubly sparse approach for group variable selection. *Annals of the Institute of Statistical Mathematics*, *69*, 997–1025. https://doi.org/10.1007/s10463-016-0571-z

Lee, S., Lee, Y., & Pawitan, Y. (2018). Sparse pathway-based prediction models for high-throughput molecular data. *Computational Statistics & Data Analysis*, *126*, 125–135. https://doi.org/10.1016/j.csda.2018.04.012

Liu, J., Huang, J., & Ma, S. (2013). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in Medicine*, *32*, 3509–3521. https://doi.org/10.1002/sim.5780

Mameli, V., Slanzi, D., Poli, I., & Green, D. V. S. (2021). Search for relevant subsets of binary predictors in high dimensional regression for discovering the lead molecule. *Pharmaceutical Statistics*, *20*, 898–915. https://doi.org/10.1002/pst.2117

Matsui, H. (2015). Sparse regularization for bi-level variable selection. *Journal of the Japanese Society of Computational Statistics*, *28*, 83–103. https://doi.org/10.5183/jjscs.1502001_216

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*, 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Qiu, C., Gelaye, B., Denis, M., Tadesse, M. G., Luque Fernandez, M. A., Enquobahrie, D. A., Ananth, C. V., Sanchez, S. E., & Williams, M. A. (2015). Circadian clock-related genetic risk scores and risk of placental abruption. *Placenta*, *36*, 1480–1486. https://doi.org/10.1016/j.placenta.2015.10.005

Rodriguez-Linares, L., Vila, X., Lado, M. J., Mendez, A., Otero, A., & Garcia, C. A. (2020). RHRV: Heart Rate Variability Analysis of ECG Data. R package version 4.2.6.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shen, L., Tang, Y., & Tang, L. C. (2021). Understanding key factors affecting power systems resilience. *Reliability Engineering System Safety*, *212*, 107621. https://doi.org/10.1016/j.ress.2021.107621

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group Lasso. *Journal of Computational Graphical Statistics*, *22*, 231–245. https://doi.org/10.1080/10618600.2012.681250

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., & Simon, M. N. (2018). *SGL: Fit a GLM (or Cox model) with a combination of Lasso and group Lasso regularization*. R package version 1.3. https://CRAN.R-project.org/package=SGL

Soret, P., Avalos, M., Wittkop, L., Commenges, D., & Thiébaut, R. (2018). Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. *BMC Medical Research Methodology*, *18*, 1–13. https://doi.org/10.1186/s12874-018-0609-4

Subrahmanya, N., & Shin, Y. C. (2010). Sparse multiple kernel learning for signal processing applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 788–798. https://doi.org/10.1109/tpami.2009.98

Taylor, J., & Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, *46*, 41–61. https://doi.org/10.1002/cjs.11313

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267–288.

Wang, X., Wang, H., Wang, S., & Yuan, J. (2021). Convex clustering method for compositional data via sparse group Lasso. *Neurocomputing*, *425*, 23–36. https://doi.org/10.1016/j.neucom.2020.10.105

Wei, F., & Zhu, H. (2012). Group coordinates descent algorithms for nonconvex penalized regression. *Computational Statistics Data Analysis*, *56*, 316–326. https://doi.org/10.1016/j.csda.2011.08.007

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., & Ma, S. (2019). A selective review of multi-level -omics data integration using variable selection. *High-Throughput*, *8*, E4. https://doi.org/10.3390/ht8010004

Xiang, S., Tong, X., & Ye, J. (2013). Efficient sparse group feature selection via nonconvex optimization. *International Conference on Machine Learning (PMLR)*, *28*(1), 284–292.

Xu, X., & Ghosh, M. (2015). Bayesian variable selection and estimation for group Lasso. *Bayesian Analysis*, *10*, 909–936. https://doi.org/10.1214/14-ba929

Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Colvin, M. M., Drazner, M. H., Filippatos, G. S., Fonarow, G. C., Givertz, M. M., Hollenberg, S. M., Lindenfeld, J., Masoudi, F. A., Mcbride, P. E., Peterson, P. N., Stevenson, L. W., & Westlake, C. (2017). ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *Journal of the American College of Cardiology*, *70*, 776–803. https://doi.org/10.1016/j.jacc.2017.04.025

Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Drazner, M. H., Fonarow, G. C., Geraci, S. A., Horwich, T., Januzzi, J. L., Johnson, M. R., Kasper, E. K., Levy, W. C., Masoudi, F. A., Mcbride, P. E., Mcmurray, J. J. V., Mitchell, J. E., Peterson, P. N., Riegel, B., … Wilkoff, B. L. (2013). ACCF/AHA guideline for the management of heart failure: Executive summary: A report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation*, *128*, 1810–1852. https://doi.org/10.1161/CIR.0b013e31829e8807

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*, 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zaharieva, M., Breiteneder, C., & Hudec, M. (2017). Unsupervised group feature selection for media classification. *International Journal of Multimedia Information Retrieval*, *6*, 233–249. https://doi.org/10.1007/s13735-017-0126-y

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942. https://doi.org/10.1214/09-AOS729

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., & Wild, P. S. (2024). Interpretability of bi-level variable selection methods. *Biometrical Journal*, *66*, 2300063. https://doi.org/10.1002/bimj.202300063