

RESEARCH ARTICLE OPEN ACCESS

A Multi-Task Learning Approach for Identifying Prognostic Variables Across Multiple Historical Clinical Trials

Xueying Liu¹  | Huijing Jiang² | Jennie Zhang² | Xinwei Deng³ 

¹Department of Statistical Science, Baylor University, Waco, Texas, USA | ²Data and Quantitative Sciences, Bristol Myers Squibb, Madison, New Jersey, USA |

³Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

Correspondence: Xinwei Deng (xdeng@vt.edu)

Received: 20 April 2026 | **Revised:** 1 September 2026 | **Accepted:** 1 October 2026

Keywords: backward selection algorithm | multi-task learning | penalized likelihood | prognostic covariates | variable selection

ABSTRACT

In clinical trials, it is critical to incorporate prognostic variables to improve precision for estimating treatment effects. Per FDA guidance, these prognostic variables need to be prospectively specified in the statistical analysis plan. Therefore, there is a great need to effectively identify prognostic variables from historical studies that are similar to the new study. In this work, we propose a multi-task learning method to identify prognostic variables from multiple related historical clinical trials. Specifically, a *bi-level constraint variable selection (Bi-CVS)* method is developed to identify both study-specific (within-study) and program-level (cross-study) prognostic variables. In addition, we develop an efficient algorithm for the *Bi-CVS*, tailored for clinical trial settings, and investigate its consistency property. The performance of the proposed method is evaluated through simulations for various endpoints, including continuous, binary, and time-to-event. A real clinical application is also presented to illustrate the proposed method.

1 | Introduction

Prognostic variables are baseline covariates that are likely to be associated with the primary endpoint of a clinical trial. They are typically demographic factors, disease characteristics, or any other clinical information relevant to each participant. Per FDA's guidance [1], in the statistical analysis of randomized clinical trials, it is critical to adjust for prognostic baseline covariates to improve precision in estimating treatment effects. Moreover, these prognostic variables and covariates adjusted analysis are required to be prospectively specified in the statistical analysis plan (SAP).

In the conduct of a clinical trial, prognostic variables to be adjusted in the primary analysis are often considered as stratification factors in patient randomization. Therefore, they need to be pre-specified in the protocol at the study design stage. However,

as no study data is yet available at the time of study design, clinical teams mostly rely on their medical knowledge and prior experience to determine the stratification factors. To guide clinical teams in their decision-making, in this paper, we propose an approach to quantitatively identify prognostic variables by leveraging data from historical studies that are similar to the new study under the design.

Several challenges arise in effectively identifying prognostic variables from multiple historical studies. First of all, due to individual study design or patient population, baseline covariates could be collected differently in each study, and category levels of factors could vary across studies as well. For instance, a regional Phase 2 study may only enroll the Asian population, whereas a global Phase 3 study could cover multiple ethnicity/race categories. Second, the impact of a baseline covariate on the primary endpoint is not expected to be the same for all studies.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

A variable could be prognostic in some of the studies but not in others. Lastly, the degree of similarity between each historical study and the new study differs, impacting their influence in determining the overall prognostic variables. The proposed approach addresses these challenges by accommodating different collections of covariates and impacts of prognostic variables across studies, as well as varying degrees of similarity between each historical study and the new study.

Because prognostic variables are closely associated with the study outcome, identifying them among demographic factors and baseline disease characteristics can be considered a variable selection problem [2]. To perform variable selection across multiple datasets, a heuristic approach is to select variables in each dataset individually and then combine these selected variable sets via some metrics. However, by pooling these historical studies together, we assume there is common information shared among these datasets that could be learned jointly. In contrast, learning individual variable selection models separately fails to leverage this shared information.

To leverage multiple related data sources, we consider a multi-task learning (MTL) approach for joint variable selection across datasets, thereby improving the general selection accuracy [3, 4]. Existing MTL methods for feature selection [5, 6] often select variables in an “all-in-all-out” fashion, identifying them as either prognostic across all studies or non-prognostic in all. This overlooks the variability in the impact of prognostic variables across different studies. To overcome this limitation, a multi-level Lasso [7], also known as meta-Lasso [8], was proposed to borrow strength across multiple datasets while keeping selection flexibility in each individual dataset. However, applying these existing methods to clinical trial studies presents several challenges. First, they assume that all datasets collect the same set of demographic and baseline disease characteristics. Second, these methods determine the number of prognostic variables to select by searching through the entire regularization path, while clinical teams only consider a very limited number (i.e., 2–4) in practice. Therefore, there is a need to decide the specified number of prognostic variables in a more straightforward and efficient way.

In this paper, we propose a *bi-level constraint variable selection (Bi-CVS)* approach to conduct multi-task prognostic variable identification across multiple historical clinical studies. The proposed method aims to select prognostic variables at both cross-study and within-study levels. The key contributions of this work are as follows.

First, we propose a bi-level constraint method to identify both cross-study and within-study prognostic variables. At the cross-study level, the proposed method uses the L_0 -norm to constrain the number of selected prognostic variables. At the within-study level, we impose an L_1 penalty on the negative log-likelihood to perform variable selection independently in each study. These study-specific penalties allow variables identified as prognostic overall to be non-prognostic in individual studies, thereby accommodating variations in their impact and enabling different covariate sets and factor levels across studies. Second, by adding an L_0 -norm constraint at the cross-study level, the proposed approach can explicitly control the number of prognostic variables to select. In clinical trial design, typically

only a few prognostic variables are considered [9–11]. Unlike traditional methods (e.g., L_1 regularization) that require tuning penalty parameters and searching through the regularization path, the proposed method directly provides the exact number of variables needed, making it easier for clinical teams to use and interpret. Moreover, it can alleviate the interdependence issue of penalty parameters for two-level sparsity problems. Finally, we develop an efficient backward selection algorithm for the *Bi-CVS* and investigate its consistency properties.

The rest of this article is organized as follows. After a brief literature review in Section 2, we detail the proposed *Bi-CVS* approach for likelihood-based models and introduce an efficient algorithm for it in Section 3. Section 4 demonstrates the performance of the proposed method via several numerical examples. A real application to the luspatercept studies is analyzed in Section 5. Finally, Section 6 concludes our work with further discussion.

2 | Review of Multi-Task Learning Feature Selection

Several techniques for variable selection in multiple datasets have been studied in the literature. Consider N datasets from related studies for variable selection, denoted as $\{S_1, \dots, S_N\}$. For study S_i , let n_i be the number of observations and p_i be the number of covariates. Denote $\mathbf{y}^{(S_i)} \in \mathbb{R}^{n_i}$ as the response vector, $\mathbf{X}^{(S_i)} \in \mathbb{R}^{n_i \times p_i}$ as the design matrix, and $\boldsymbol{\beta}^{(S_i)} = (\beta_1^{(S_i)}, \dots, \beta_{p_i}^{(S_i)})^T$ as the coefficient vector. To find a common set of variables that significantly affect the response across different datasets, a heuristic approach is to separately maximize N penalized log-likelihoods (or minimize its negative form), each of which is formulated as

$$\ell(\boldsymbol{\beta}^{(S_i)}) - \lambda_i \|\boldsymbol{\beta}^{(S_i)}\|_1, \quad i = 1, \dots, N,$$

where $\ell(\boldsymbol{\beta}^{(S_i)})$ is the log-likelihood, λ_i is the non-negative penalty parameter, and L_1 regularization $\|\boldsymbol{\beta}^{(S_i)}\|_1 = \sum_{k=1}^{p_i} |\beta_k^{(S_i)}|$ is the Lasso penalty that shrinks some coefficients $\beta_k^{(S_i)}$ to exactly 0 [12], selecting only a subset of variables. The Lasso can be extended to group Lasso to deal with the problem of selecting grouped variables [13]. Upon selecting variables independently within each task, feature selection across studies is achieved by combining these selected variable sets via some metrics, for example, the majority vote. However, this approach overlooks the shared information across these related datasets, as it conducts variable selection separately in each dataset.

To leverage shared information and improve general performance, multi-task learning (MTL) methods have been proposed where multiple related tasks are learned jointly [3]. Assuming that all datasets have the same set of covariates, that is, $p_1 = \dots = p_N = p$, the multi-task feature selection problem [5, 6] is formulated through $L_{2,1}$ regularization for variables of different datasets as follows:

$$\min_{\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(N)}} - \sum_{i=1}^N \ell(\boldsymbol{\beta}^{(S_i)}) + \lambda \sum_{k=1}^p \|\boldsymbol{\beta}_k\|_2,$$

where $\boldsymbol{\beta}_k = (\beta_k^{(S_1)}, \dots, \beta_k^{(S_N)})^T$ is the coefficient vector of the k th variable across N datasets. This approach penalizes the

L_1 -norm of the vector $(\|\beta_1\|_2, \dots, \|\beta_p\|_2)^T$, pushing several elements $\|\beta_k\|_2 = \left(\sum_{i=1}^N (\beta_k^{(S_i)})^2\right)^{1/2}$ to be 0. This implies $\beta_k^{(S_i)} = 0$ for all study S_i , removing variable x_k from all datasets. This $L_{2,1}$ regularization extends the L_1 regularization by summing the L_2 -norms of grouped coefficients. Its main limitation is that it selects variables in an “all-in-all-out” fashion, either selecting a variable in all datasets or excluding it from all. This lack of flexibility limits its applicability to real-world problems where the impact of variables varies across datasets [14].

To overcome this challenge, a multi-level Lasso [7], also called meta-Lasso [8], was proposed to allow a variable to be selected at cross-study level but can be unimportant in specific studies. In this method, the coefficient $\beta_k^{(S_i)}$ is decomposed into a global component $\theta_k \geq 0$ and a study-specific component $\gamma_k^{(S_i)}$, that is, $\beta_k^{(S_i)} = \theta_k \gamma_k^{(S_i)}$, $k = 1, \dots, p$, $i = 1, \dots, N$. Then two penalty parameters are involved: λ_1 is to control variable selection at the cross-study level, and λ_2 is for the individual-study level:

$$\min_{\theta, \gamma} - \sum_{i=1}^N \ell(\theta, \gamma^{(S_i)}) + \lambda_1 \sum_{k=1}^p |\theta_k| + \lambda_2 \sum_{k=1}^p \sum_{i=1}^N |\gamma_k^{(S_i)}|,$$

where $\theta = (\theta_1, \dots, \theta_p)^T$ and $\gamma^{(S_i)} = (\gamma_1^{(S_i)}, \dots, \gamma_p^{(S_i)})^T$. If θ_k is shrunk to 0, then $\beta_k^{(S_i)}$ will become 0 for $i = 1, \dots, N$, implying that the variable x_k is excluded from all datasets. Therefore, the global components θ_k control the sparsity across studies, while $\gamma_k^{(S_i)}$ controls sparsity within study S_i .

However, these existing MTL methods assume the same set of covariates across studies, limiting their applicability in modeling multiple clinical trials with varying study designs or patient populations. Additionally, they require searching through the regularization path to determine the number of variables requested by clinical teams, which increases computational complexity. To overcome these limitations and address the challenges outlined in Section 1, we propose a novel *bi-level constraint variable selection* (Bi-CVS) approach that accommodates diverse covariate sets, varying impacts of prognostic variables, and different degrees of similarity across studies. Moreover, it can explicitly control the number of prognostic variables to be selected, enhancing efficiency and interpretability.

3 | Proposed Method

In this section, we describe the *bi-level constraint variable selection* (Bi-CVS) approach within the multi-task learning framework. Additionally, we present a backward selection algorithm for the Bi-CVS and study its consistency.

3.1 | Multi-Task Learning Across Multiple Historical Clinical Trials

We consider N historical clinical trials, denoted by $S_1, \dots, S_N$, that are similar to the new study. Let $\mathcal{P}_0 = \{1, \dots, p\}$ index the

p unique variables pooled from these trials, and let x_k denote the k th pooled variable. These variables may fully overlap across studies if they are collected in every trial, or partially overlap if they are collected only in a subset of trials.

For each study S_i , let $\mathcal{V}_i \subseteq \mathcal{P}_0$ denote the index set of variables collected in that study. The corresponding design matrix is written as $\mathbf{X}^{(i)} = [\mathbf{x}_k^{(i)} : k \in \mathcal{V}_i]$, where $\mathbf{x}_k^{(i)}$ denotes the observed data vector of variable x_k in study S_i . Similarly, let $\beta_k^{(i)}$ denote the coefficient of variable x_k in study S_i , and let $\beta^{(i)} = (\beta_k^{(i)} : k \in \mathcal{V}_i)^T$ be the corresponding coefficient vector. Let $\mathbf{A}^{(i)}$ be the treatment indicator vector, $\tau^{(i)}$ the treatment effect, and $\mathbf{y}^{(i)}$ the response vector for study S_i . Assuming independence across studies, the joint likelihood factorizes as

$$L(\beta^{(1)}, \tau^{(1)}, \dots, \beta^{(N)}, \tau^{(N)}) = \prod_{i=1}^N L_i(\beta^{(i)}, \tau^{(i)} | \mathbf{A}^{(i)}, \mathbf{X}^{(i)}, \mathbf{y}^{(i)}).$$

For two-arm studies, $A^{(i)}$ is the usual binary treatment indicator. For single-arm studies, we set $A^{(i)} \equiv 0$ for all subjects. In this case, the treatment effect parameter is not separately identifiable and can be omitted from the optimization without affecting the proposed cross-study selection procedure. Since the treatment effect is not penalized, its omission does not introduce any additional scaling or penalization of the baseline covariates.

To model the studies jointly, we minimize the weighted negative log-likelihood together with a penalty $\phi(\beta)$ on the cross-study coefficient vector $\beta = ((\beta^{(1)})^T, \dots, (\beta^{(N)})^T)^T$:

$$-\ell(\beta, \tau) + \phi(\beta) = - \sum_{i=1}^N w_i \ell_i(\beta^{(i)}, \tau^{(i)}) + \phi(\beta),$$

where ℓ_i is the log-likelihood for study S_i and w_i is its study-specific weight. The weights are nonnegative and normalized such that $\sum_{i=1}^N w_i = 1$. Thus, their relative magnitudes determine the contributions of the historical studies to the joint objective. By introducing the penalty term $\phi(\beta)$, we impose cross-study structure on the coefficients and thereby enable multi-task learning. This framework can be used with a variety of endpoints, including linear regression for continuous outcomes, logistic regression for binary outcomes, and the Cox proportional hazards model for time-to-event outcomes.

The weighting scheme can be used for incorporating differences in the relevance of historical studies and does not require a single weighting rule for all applications. In pharmaceutical practice, the weights may be prespecified in consultation with clinical and other domain experts based on their knowledge of the historical and new studies. One possible approach is to compare the patient enrollment criteria, including both inclusion and exclusion criteria, of each historical study with those of the new study. For example, let c_i denote the number of criteria shared by study S_i and the new study. The weight may then be defined as $w_i = c_i / \sum_{j=1}^N c_j$, so that studies sharing more enrollment criteria with the new study receive greater influence in the joint selection procedure. This overlap-based measure is one possible implementation of the weighting framework; other prespecified

measures may be used when supported by the clinical team. If no reliable prior information is available to distinguish the relevance of the historical studies, one can use equal weights $w_i = 1/N$ as a default.

3.2 | Bi-Level Constraint Variable Selection (Bi-CVS)

Given the multi-task learning framework for multiple historical studies, the next step is to identify prognostic variables across them. This is a variable selection problem aimed at identifying the variables most strongly associated with the outcome. Designed for clinical trial data, we propose a method that selects a pre-specified number of prognostic variables while accommodating different variable collections, categorizations, prognostic effects, and degrees of similarity across studies.

For each pooled variable x_k , let $\mathcal{I}_k = \{i : k \in \mathcal{V}_i\}$ denote the set of studies in which x_k is collected. Define $\tilde{\beta}_k = (\beta_k^{(i)} : i \in \mathcal{I}_k)^T$, and collect the corresponding group norms into $\tilde{\mathbf{g}} = (\|\tilde{\beta}_1\|_2, \dots, \|\tilde{\beta}_p\|_2)^T$.

At the first level, our goal is to identify cross-study prognostic variables. To enforce sparsity among the cross-study groups $\tilde{\beta}_k$, $k = 1, \dots, p$, we impose an L_0 -constraint on $\tilde{\mathbf{g}}$:

$$\|\tilde{\mathbf{g}}\|_0 = \sum_{k=1}^p I\{\|\tilde{\beta}_k\|_2 \neq 0\} \leq \eta.$$

This constraint limits the number of pooled variables that remain prognostic across studies to at most η . If $\|\tilde{\beta}_k\|_2 = 0$, then $\beta_k^{(i)} = 0$ for all $i \in \mathcal{I}_k$, indicating that x_k is non-prognostic in every study where it is collected.

At the second level, we allow the impact of a cross-study prognostic variable to vary from study to study. If x_k is categorical, its coefficient in study S_i is represented by a group of dummy-variable coefficients $\{\beta_{k,l}^{(i)} : l = 1, \dots, l_k^{(i)}\}$, where $l_k^{(i)}$ denotes number of non-reference dummy coefficients of x_k in study S_i . This allows the same categorical variable to have different numbers of levels across studies. Within each study, we use a group Lasso penalty

so that all levels of a categorical variable are selected or excluded together:

$$(\hat{\beta}^{(i)}, \hat{\tau}^{(i)}) = \arg \min_{\beta^{(i)}, \tau^{(i)}} \left\{ -\frac{1}{n_i} \ell_i(\beta^{(i)}, \tau^{(i)}) + \lambda_i \sum_{k \in \mathcal{V}_i} \sqrt{l_k^{(i)}} \|\beta_k^{(i)}\|_2 \right\}, \quad (1)$$

where λ_i is a nonnegative penalty parameter that can be selected by cross-validation. For a categorical variable x_k , $\|\beta_k^{(i)}\|_2 = \left(\sum_{l=1}^{l_k^{(i)}} (\beta_{k,l}^{(i)})^2 \right)^{1/2}$. For a continuous variable x_k , the term $\sqrt{l_k^{(i)}} \|\beta_k^{(i)}\|_2$ reduces to $|\beta_k^{(i)}|$, which corresponds to a Lasso penalty.

Combining the constraints from these two levels, we define the proposed *bi-level constraint variable selection* (Bi-CVS) estimator

$$(\hat{\beta}^{(1)}, \hat{\tau}^{(1)}, \dots, \hat{\beta}^{(N)}, \hat{\tau}^{(N)}),$$

as the solution to

$$\arg \min_{\substack{\beta^{(i)}, \tau^{(i)} \\ i=1, \dots, N}} \sum_{i=1}^N w_i \left\{ -\frac{1}{n_i} \ell_i(\beta^{(i)}, \tau^{(i)}) + \lambda_i \sum_{k \in \mathcal{V}_i} \sqrt{l_k^{(i)}} \|\beta_k^{(i)}\|_2 \right\} \\ \text{s.t. } \|\tilde{\mathbf{g}}\|_0 = \sum_{k=1}^p I\{\|\tilde{\beta}_k\|_2 \neq 0\} \leq \eta. \quad (2)$$

Here, w_i is the weight for study S_i defined in Section 3.1, λ_i is the study-specific penalty parameter selected by cross-validation, and η specifies the desired number of prognostic variables across the N studies. As illustrated in Figure 1, the first level constraint enforces sparsity across studies, removing non-prognostic variables, while the second level constraint accommodates study-specific variations in the impact of selected variables.

In clinical trial applications, it is beneficial for researchers to be able to specify the exact number of prognostic variables they need. This ensures that the results are interpretable and aligned with domain-specific knowledge. The proposed method, utilizing the L_0 -norm, provides explicit control over the number of required prognostic variables across studies through the parameter η . If the optimal number of prognostic variables is unknown, η can be treated as a tuning parameter and determined via cross-validation [15].

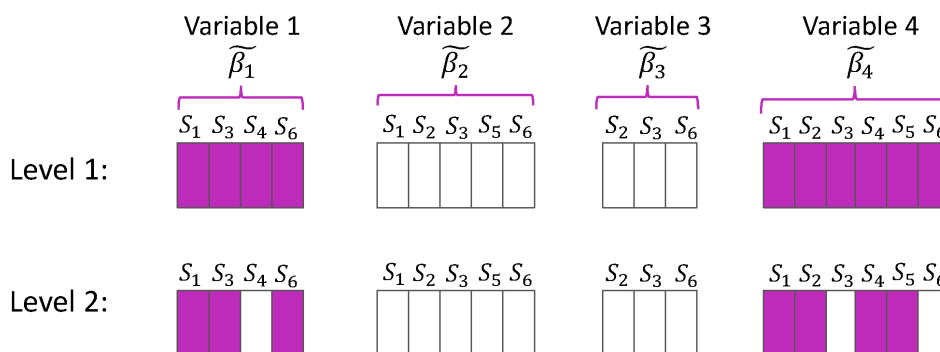


FIGURE 1 | Illustration of the two-level sparsity in the Bi-CVS. Each box represents a variable in a study, and grouped boxes represent identical variables across studies. Pink boxes denote variables identified as prognostic. At Level 1, $\eta = 2$, Variables 1 and 4 are identified as prognostic across all studies that include them. Then Level 2 sparsity allows these variables to be non-prognostic in specific studies, such as Variable 1 in study 4.

3.3 | Backward Selection Algorithm

Although L_0 regularization is ideal for enforcing sparsity by consistently finding the best η -sparse estimator [16, 17], its NP-hard nature poses computational challenges, as the number of candidate subsets grows exponentially with p [18]. Therefore, stepwise selection methods are commonly used in practice, including forward selection, backward selection, and their combination, each with its own advantages and limitations [19, 20].

In clinical trial settings, increasing the number of prognostic variables adds complexity to implementing stratified randomization and may lead to imbalanced distributions across an excessive number of strata [21]. Since only a limited number of prognostic factors (typically 2–4) are used for stratification in practice, backward selection is particularly suitable. Unlike forward selection, which may miss important variables because of its limited selection steps, backward selection starts from the full pooled set and sequentially removes the least influential variables. In addition, clinical trial datasets typically satisfy $n > p$, so backward selection can be applied without the severe overfitting concerns that arise in high-dimensional settings. Finally, backward selection is simple and computationally efficient compared with stepwise procedures and other penalty alternatives (e.g., L_1 -norm [12], L_2 -norm [22], and their combinations [23]) that require going through the regularization path to find the penalty parameter yielding exactly η non-zero variables. We therefore adopt a backward selection algorithm for the proposed *Bi-CVS* model.

Given input datasets $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{X}^{(N)}, \mathbf{y}^{(N)})$, the algorithm is initialized with the full pooled index set $\mathcal{P}_0 = \{1, \dots, p\}$. At each iteration, let $\mathcal{P} \subseteq \mathcal{P}_0$ denote the current candidate set. For each study S_i , Equation (1) is fitted using the variables indexed by $\mathcal{P}_i = \mathcal{P} \cap \mathcal{V}_i$, that is, the variables currently under consideration and observed in study S_i . The study-specific tuning parameter is selected by cross-validation based on the current variable set $\mathcal{P}$ and is denoted by $\hat{\lambda}_i(\mathcal{P})$. These tuning parameters are then held fixed when evaluating all candidate variable removals at the current backward-selection step. The empirical objective value for $\mathcal{P}$, denoted by $Q_n(\mathcal{P})$, is then evaluated according to Equation (2). For each $k \in \mathcal{P}$, the PMLE models are refitted using $\mathcal{P} \setminus \{k\}$ with the current $\hat{\lambda}_i(\mathcal{P})$, and the effect of removing variable index k is measured by $D_n(k; \mathcal{P}) = Q_n(\mathcal{P} \setminus \{k\}) - Q_n(\mathcal{P})$. The variable index corresponding to the smallest increase in the objective value is removed from $\mathcal{P}$. After removal, the study-specific tuning parameters are re-selected by cross-validation for the affected studies. Fixing $\hat{\lambda}_i(\mathcal{P})$ during candidate evaluation avoids repeated cross-validation and reduces computation when p is large. This backward elimination procedure continues until $|\mathcal{P}| = \eta$, at which point we set $\hat{\mathcal{P}} = \mathcal{P}$.

Suppose that q variables remain at a given backward-elimination step. There are q candidate removals, and, with the study-specific $\hat{\lambda}_i$ held fixed, each candidate may require refitting the penalized model in all N studies. Assuming a bounded number of optimization iterations and a bounded number of design columns per pooled variable, a study-specific fit with n_i observations has complexity $O(n_i q)$. Thus, candidate evaluation over the full elimination path has complexity $O\left\{\left(\sum_{i=1}^N n_i\right)(p^3 - \eta^3)\right\}$.

ALGORITHM 1 | Backward selection algorithm.

Require: Datasets $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{X}^{(N)}, \mathbf{y}^{(N)})$; target size η

Ensure: Selected cross-study prognostic index set $\hat{\mathcal{P}}$

- 1: **Initialization:**
- 2: Set $\mathcal{P} \leftarrow \mathcal{P}_0 = \{1, \dots, p\}$
- 3: For each study S_i , define $\mathcal{P}_i = \mathcal{P} \cap \mathcal{V}_i$
- 4: Fit the PMLE model in Equation~(2) using the variables indexed by $\mathcal{P}_i$ and determine $\hat{\lambda}_i$ by cross-validation
- 5: Compute $Q_n(\mathcal{P})$ according to Equation~(2)
- 6: **Backward Selection:**
- 7: **while** $|\mathcal{P}| > \eta$ **do**
- 8: **for** each $k \in \mathcal{P}$ **do**
- 9: Define $\mathcal{P}^{-k} = \mathcal{P} \setminus \{k\}$
- 10: For each study S_i , define $\mathcal{P}_i^{-k} = \mathcal{P}^{-k} \cap \mathcal{V}_i$
- 11: Refit the PMLE model using $\mathcal{P}_i^{-k}$ with the current $\hat{\lambda}_i$ and compute $Q_n(\mathcal{P}^{-k})$
- 12: Compute $D_n(k; \mathcal{P}) = Q_n(\mathcal{P}^{-k}) - Q_n(\mathcal{P})$
- 13: **end for**
- 14: Select $k^* \in \arg \min_{k \in \mathcal{P}} D_n(k; \mathcal{P})$
- 15: Update $\mathcal{P} \leftarrow \mathcal{P} \setminus \{k^*\}$
- 16: Re-select $\hat{\lambda}_i$ by cross-validation and update $Q_n(\mathcal{P})$
- 17: **end while**
- 18: Set $\hat{\mathcal{P}} = \mathcal{P}$

After a variable is selected, cross-validation is repeated only for the affected studies. With K folds and L tuning parameter values along the regularization path, this contributes $O\left\{KL\left(\sum_{i=1}^N n_i\right)(p^2 - \eta^2)\right\}$. Therefore, when $n_i = O(n)$, the overall worst-case complexity is

$$O[Nn\{(p^3 - \eta^3) + KL(p^2 - \eta^2)\}].$$

In practice, parallel computing techniques can be used to evaluate candidate variable removals simultaneously, thereby reducing the overall computing time.

3.4 | Consistency Properties

Algorithm 1 initializes with the full pooled variable set $\mathcal{P}_0$ and iteratively removes the variable yielding the smallest empirical deletion increment until exactly $|\mathcal{P}| = \eta$ variables remain. In this section, we study the large-sample behavior of this backward elimination procedure under a fixed-dimensional framework, where both the number of studies N and the pooled dimension p remain fixed as the study sample sizes n_i increase. Specifically, we show that the backward elimination path based on the empirical objective converges to the corresponding deterministic path defined by its population objective.

To reflect the implementation of Algorithm 1, the tuning parameter is allowed to depend on the candidate subset under consideration. For each study i and each candidate set $\mathcal{P} \subseteq \mathcal{P}_0$, let $\hat{\lambda}_i(\mathcal{P})$ denote the study-specific tuning parameter selected when fitting the penalized model on the subset $\mathcal{P}_i = \mathcal{P} \cap \mathcal{V}_i$.

For any candidate set $\mathcal{P} \subseteq \mathcal{P}_0$, define the empirical objective

$$Q_n(\mathcal{P}) = \sum_{i=1}^N w_i \inf_{\beta_{\mathcal{P}}^{(i)}, \tau^{(i)}} \left\{ -\frac{1}{n_i} \ell_i(\beta_{\mathcal{P}}^{(i)}, \tau^{(i)}) + \hat{\lambda}_i(\mathcal{P}) \sum_{k \in \mathcal{P}_i} \sqrt{l_k^{(i)}} \|\beta_k^{(i)}\|_2 \right\}.$$

That is, $Q_n(\mathcal{P})$ evaluates each subset at its own tuned penalty level. Let $\lambda_i(\mathcal{P})$ denote the deterministic large-sample limit of $\hat{\lambda}_i(\mathcal{P})$. We define the corresponding population objective by

$$Q(\mathcal{P}) = \sum_{i=1}^N w_i \inf_{\beta_{\mathcal{P}}^{(i)}, \tau^{(i)}} \left\{ E \left[-\frac{1}{n_i} \ell_i(\beta_{\mathcal{P}}^{(i)}, \tau^{(i)}) \right] + \lambda_i(\mathcal{P}) \sum_{k \in \mathcal{P}_i} \sqrt{l_k^{(i)}} \|\beta_k^{(i)}\|_2 \right\}.$$

For any $k \in \mathcal{P}$, define the empirical and population deletion increments by $D_n(k; \mathcal{P}) = Q_n(\mathcal{P} \setminus \{k\}) - Q_n(\mathcal{P})$ and $D(k; \mathcal{P}) = Q(\mathcal{P} \setminus \{k\}) - Q(\mathcal{P})$. Starting from $\mathcal{P}_{\text{pop}}^{(0)} = \mathcal{P}_0$, the population backward elimination path is defined by repeatedly removing $k_t^* = \arg \min_{k \in \mathcal{P}_{\text{pop}}^{(t)}} D(k; \mathcal{P}_{\text{pop}}^{(t)})$, and setting $\mathcal{P}_{\text{pop}}^{(t+1)} = \mathcal{P}_{\text{pop}}^{(t)} \setminus \{k_t^*\}$, $t = 0, 1, \dots, p - \eta - 1$. The final population-selected set is denoted as $\mathcal{P}_{\text{pop}} = \mathcal{P}_{\text{pop}}^{(p-\eta)}$.

The assumptions below are standard for fixed- p penalized M -estimation over a finite collection of candidate subsets. Conditions (C1) and (C2) are routine regularity conditions for the loss and parameter spaces. Condition (C3) formalizes the asymptotic stability of the subset-specific tuning parameters, while Condition (C4) is a separation condition imposed to ensure that the variable deleted at each population step is uniquely determined.

(C1) The number of studies N and the pooled dimension p are fixed, and $\min_{1 \leq i \leq N} n_i \rightarrow \infty$.

(C2) For each study i and each subset $\mathcal{P} \subseteq \mathcal{P}_0$, the parameter space corresponding to $(\beta_{\mathcal{P}}^{(i)}, \tau^{(i)})$ is compact, the negative log-likelihood function and the penalty term are continuous in the parameters, the negative log-likelihood is dominated by an integrable envelope, and the infima defining $Q_n(\mathcal{P})$ and $Q(\mathcal{P})$ are attained.

(C3) For each study $i = 1, \dots, N$, $\sup_{\mathcal{P} \subseteq \mathcal{P}_0} |\hat{\lambda}_i(\mathcal{P}) - \lambda_i(\mathcal{P})| \rightarrow 0$.

(C4) Along the population backward elimination path, the variable to be removed at each step is unique. Moreover, there exists a constant $\Delta_{\min} > 0$ such that for every $t = 0, 1, \dots, p - \eta - 1$,

$$\min_{k \in \mathcal{P}_{\text{pop}}^{(t)}, k \neq k_t^*} \left\{ D(k; \mathcal{P}_{\text{pop}}^{(t)}) - D(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) \right\} \geq \Delta_{\min}.$$

We first establish a uniform convergence result for $Q_n(\mathcal{P})$. Its proof is based on the uniform convergence of the empirical loss to its population counterpart in Condition (C2), the asymptotic stability of the subset-specific tuning parameters in Condition (C3), and the finiteness of the candidate subset class under Condition (C1). The proof is deferred to Appendix A in [Supporting Information](#). With Lemma 1, the proof of path consistency reduces to comparing the empirical and population deletion increments.

Lemma 1. Under Conditions (C1)–(C3),

$$\sup_{\mathcal{A} \subseteq \mathcal{P}_0} |Q_n(\mathcal{A}) - Q(\mathcal{A})| \xrightarrow{p \rightarrow 0} 0 \quad \text{as } \min_{1 \leq i \leq N} n_i \rightarrow \infty.$$

Theorem 1. (Consistency of subset-wise tuned backward elimination). Let $\hat{\mathcal{P}}$ be the output of Algorithm 1 when the algorithm stops at size η , and let $\mathcal{P}_{\text{pop}}$ be the corresponding population result defined above. Under Conditions (C1)–(C4),

$$\Pr(\hat{\mathcal{P}} = \mathcal{P}_{\text{pop}}) \rightarrow 1 \quad \text{as } \min_{1 \leq i \leq N} n_i \rightarrow \infty.$$

Proof. For any candidate set $\mathcal{P} \subseteq \mathcal{P}_0$ and any $k \in \mathcal{P}$, recall that

$$D_n(k; \mathcal{P}) = Q_n(\mathcal{P} \setminus \{k\}) - Q_n(\mathcal{P}), \quad D(k; \mathcal{P}) = Q(\mathcal{P} \setminus \{k\}) - Q(\mathcal{P}).$$

By Lemma 1,

$$\sup_{\mathcal{A} \subseteq \mathcal{P}_0} |Q_n(\mathcal{A}) - Q(\mathcal{A})| \xrightarrow{p \rightarrow 0} 0.$$

Hence, for any $\mathcal{P} \subseteq \mathcal{P}_0$ and any $k \in \mathcal{P}$,

$$\begin{aligned} |D_n(k; \mathcal{P}) - D(k; \mathcal{P})| &= \left| [Q_n(\mathcal{P} \setminus \{k\}) - Q(\mathcal{P} \setminus \{k\})] \right. \\ &\quad \left. - [Q_n(\mathcal{P}) - Q(\mathcal{P})] \right| \\ &\leq |Q_n(\mathcal{P} \setminus \{k\}) - Q(\mathcal{P} \setminus \{k\})| \\ &\quad + |Q_n(\mathcal{P}) - Q(\mathcal{P})| \\ &\leq 2 \sup_{\mathcal{A} \subseteq \mathcal{P}_0} |Q_n(\mathcal{A}) - Q(\mathcal{A})|. \end{aligned}$$

Taking the supremum overall $\mathcal{P} \subseteq \mathcal{P}_0$ and $k \in \mathcal{P}$ yields

$$\sup_{\mathcal{P} \subseteq \mathcal{P}_0} \sup_{k \in \mathcal{P}} |D_n(k; \mathcal{P}) - D(k; \mathcal{P})| \xrightarrow{p \rightarrow 0} 0.$$

Define the event

$$\mathcal{E}_n = \left\{ \sup_{\mathcal{P} \subseteq \mathcal{P}_0} \sup_{k \in \mathcal{P}} |D_n(k; \mathcal{P}) - D(k; \mathcal{P})| < \frac{\Delta_{\min}}{3} \right\}.$$

Then $\Pr(\mathcal{E}_n) \rightarrow 1$. We claim that on $\mathcal{E}_n$, the empirical and population backward elimination procedures make the same deletion at every step. Fix $t \in \{0, 1, \dots, p - \eta - 1\}$, and suppose the current population candidate set is $\mathcal{P}_{\text{pop}}^{(t)}$. By Condition (C4), the population minimizer is uniquely k_t^* , and for every $k \in \mathcal{P}_{\text{pop}}^{(t)}$ with $k \neq k_t^*$,

$$D(k; \mathcal{P}_{\text{pop}}^{(t)}) \geq D(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) + \Delta_{\min}. \quad (3)$$

On the event $\mathcal{E}_n$, for $k \in \mathcal{P}_{\text{pop}}^{(t)}$,

$$D_n(k; \mathcal{P}_{\text{pop}}^{(t)}) \geq D(k; \mathcal{P}_{\text{pop}}^{(t)}) - \frac{\Delta_{\min}}{3}, \quad (4)$$

and for $k_t^* \in \mathcal{P}_{\text{pop}}^{(t)}$,

$$D_n(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) \leq D(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) + \frac{\Delta_{\min}}{3}. \quad (5)$$

Therefore, using Equations (3–5),

$$\begin{aligned} D_n(k; \mathcal{P}_{\text{pop}}^{(t)}) - D_n(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) &\geq \left[D(k; \mathcal{P}_{\text{pop}}^{(t)}) - \frac{\Delta_{\min}}{3} \right] - \left[D(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) + \frac{\Delta_{\min}}{3} \right] \\ &\geq D(k; \mathcal{P}_{\text{pop}}^{(t)}) - D(k_t^*; \mathcal{P}_{\text{pop}}^{(t)}) - \frac{2\Delta_{\min}}{3} \end{aligned}$$

$$= D(k; \mathcal{P}_{\text{pop}}^{(t)}) - D(k_i^*; \mathcal{P}_{\text{pop}}^{(t)}) - \frac{2\Delta_{\min}}{3} \\ \geq \frac{\Delta_{\min}}{3} > 0.$$

so the empirical minimizer is also uniquely k_i^* .

Since both procedures start from the same full set $\mathcal{P}_0$, a finite induction over the $p - \eta$ elimination steps shows that they remove the same variable at every step on the event $\mathcal{E}_n$. Hence, $\widehat{\mathcal{P}} = \mathcal{P}_{\text{pop}}$ on $\mathcal{E}_n$. Because $\Pr(\mathcal{E}_n) \rightarrow 1$, it follows that $\Pr(\widehat{\mathcal{P}} = \mathcal{P}_{\text{pop}}) \rightarrow 1$. $\square$

4 | Simulation Study

In this section, we conduct simulation studies to evaluate the performance of the proposed *Bi-CVS*. We aim to demonstrate its variable selection performance across various data types and its ability to handle factors effectively. The proposed *Bi-CVS* is compared with the separate Lasso (*SepLasso*), the multi-task feature selection (*MTLLasso*), and the meta Lasso (*MetaLasso*).

4.1 | Data Generation

We first generate the study-specific covariates $\mathbf{X}$, followed by the true coefficients $\boldsymbol{\beta}$ and responses $\mathbf{y}$. We consider $N = 8$ historical studies with $n_i = 100$ and $p \in \{20, 90, 120\}$. In all settings, $\eta = 4$ pooled variables are selected as the true cross-study prognostic variables, leaving 16, 86, and 116 noise variables, respectively. We use equal study weights $w_i = 1/N$ because no prior information on study relevance is assumed.

A primary motivation for *Bi-CVS* is the heterogeneity across historical studies, where different variables may be collected or the same categorical variable may have different factor levels. Since *MTLLasso* and *MetaLasso* require a common covariate structure, we consider two scenarios:

S1. All studies contain the same set of p variables with consistent factor levels, allowing comparison of *Bi-CVS*, *SepLasso*, *MTLLasso*, and *MetaLasso* (*MTLLasso* does not support time-to-event responses).

S2. Studies may collect different subsets of the pooled variables and use different factor levels. Only *Bi-CVS* and *SepLasso* are compared in this heterogeneous setting.

In (S2), the number of variables collected in study S_i is obtained by rounding a draw from $N(0.8p, 1)$, after which the variables are randomly sampled from the pooled set. All pooled variables are required to appear in at least one study, and each true cross-study prognostic variable must be measured in at least four studies. We set 70% of the pooled variables to be categorical and the remaining variables to be continuous. Continuous variables are independently generated from $N(0, 1)$. Categorical variables have two to four levels, with at least five observations in each level. Factor levels are consistent across studies in (S1) but are generated separately within each study in (S2).

We next generate the true coefficient vector $\boldsymbol{\beta}$ by randomly selecting $\eta = 4$ pooled variables as the true cross-study prognostic variables. For each study S_i , a level-2 sparsity probability is generated from $\text{Unif}(0, 0.3)$, allowing a cross-study prognostic variable to be inactive in some studies. Each true prognostic variable is required to have a nonzero effect in at least four studies, while non-prognostic variables have coefficient zero in all studies. For an active continuous variable, $\beta_{ik} = s_{ik}a_{ik}$, where $s_{ik} \in \{-1, 1\}$ and $a_{ik} \sim \text{Unif}(0.5, 0.8)$. For an active categorical variable, category-specific effects are centered at zero and scaled to have a similar magnitude.

Then we generate three types of endpoints $\mathbf{y}$: continuous, binary, and time-to-event.

1. Continuous endpoint: For study S_i , $\mathbf{y}_i = \beta_{0i} + \mathbf{X}_i\boldsymbol{\beta}_i + \epsilon_i$, where $\beta_{0i} \sim N(0, 1)$ and $\epsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$.
2. Binary endpoint: For study S_i , we generate $\beta_{0i} \sim \text{Unif}(-0.3, 0.3)$ and calculate $\boldsymbol{\eta}_i = \beta_{0i} + \mathbf{X}_i\boldsymbol{\beta}_i$ and $p_i = \exp(\boldsymbol{\eta}_i) / (1 + \exp(\boldsymbol{\eta}_i))$. We then generate $y_{ij} \sim \text{Bernoulli}(p_{ij})$ and require at least 10 observations in each response class within each study.
3. Time-to-event endpoint: Event times are generated from a proportional hazards model with $\lambda_{ij} = \lambda_{0i} \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}_i)$, where $\lambda_{0i} \sim \text{Unif}(0.5, 1.5)$, and $T_{ij} \sim \text{Exp}(\lambda_{ij})$. Independent censoring times are generated as $C_{ij} \sim \text{Exp}(\lambda_{ci})$, where λ_{ci} is selected to achieve approximately 35% censoring within each study. The observed outcome is $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$ with event indicator $\delta_{ij} = I(T_{ij} \leq C_{ij})$. We require at least 10 observed events within each study.

4.2 | Method Comparison

We evaluate both cross-study prognostic variable selection and within-study sparsity recovery. Cross-study prognostic variable selection corresponds to Level-1 selection, while within-study sparsity recovery corresponds to Level-2 selection. For Level-1 selection, we report the exact recovery rate and recall. For each replication, exact recovery is indicated by $I(\widehat{\mathcal{P}} = \mathcal{P}^*)$, where $\mathcal{P}^*$ and $\widehat{\mathcal{P}}$ denote the true and estimated sets of cross-study prognostic variables, and the exact recovery rate is its average across simulation replications. Level-1 recall is defined as $|\widehat{\mathcal{P}} \cap \mathcal{P}^*| / |\mathcal{P}^*|$.

For Level-2 selection, each measured study-variable pair is classified as active or inactive according to whether its true and estimated coefficients are nonzero. For a categorical variable, it is considered active if at least one of its corresponding category-specific coefficients is nonzero. The following quantities are calculated over all measured study-variable pairs:

$$\text{TP} = \sum_{i,k} I(\beta_k^{(S_i)} \neq 0 \wedge \hat{\beta}_k^{(S_i)} \neq 0), \quad \text{FP} = \sum_{i,k} I(\beta_k^{(S_i)} = 0 \wedge \hat{\beta}_k^{(S_i)} \neq 0), \\ \text{TN} = \sum_{i,k} I(\beta_k^{(S_i)} = 0 \wedge \hat{\beta}_k^{(S_i)} = 0), \quad \text{FN} = \sum_{i,k} I(\beta_k^{(S_i)} \neq 0 \wedge \hat{\beta}_k^{(S_i)} = 0).$$

We report the F_1 score, false negative rate (FNR), and false positive rate (FPR), defined as $F_1 = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$, $\text{FNR} = \text{FN}/(\text{TP} + \text{FN})$, and $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$.

For a comparable Level-1 evaluation, all methods identify the same target number $\eta = 4$ of cross-study prognostic variables. *Bi-CVS* directly selects η variables through backward elimination. *SepLasso* ranks variables according to their cross-study selection frequencies, while *MTLLasso* and *MetaLasso* rank variables according to the magnitude of their estimated coefficients across studies. For these benchmark methods, the top η ranked variables are retained for Level-1 evaluation. The original study-specific estimates are used to calculate F_1 , FNR, and FPR. The comparison results for (S1) and (S2) are summarized in Tables 1 (Figure 2) and 2, respectively.

In (S1), as shown in Table 1 (Figure 2), *Bi-CVS* shows consistently strong performance across all endpoints and dimensions. For the continuous and survival endpoints, it achieves perfect Level-1 recovery for all p , with F_1 scores around 0.94–0.96 and very low FPR and FNR. For the binary endpoint, *Bi-CVS* also maintains high Level-1 exact recovery rates of 0.96–0.98, recall of at least 0.99, and F_1 scores above 0.91. Although *MTLLasso* and *MetaLasso* perform competitively in some settings, *Bi-CVS* provides more consistent performance across endpoint types and dimensions, while *SepLasso* generally produces more false-positive selections and lower F_1 scores. The improved selection performance comes with additional computational cost: the run time of *Bi-CVS* increases with p and is generally longer than that of *SepLasso* and *MTLLasso*. However, its performance remains stable as p increases from 20 to 120. Even when $p = 120 > n_i = 100$, it retains nearly perfect Level-1 recovery and low error rates, providing empirical evidence of its robustness when $p > n_i$.

In (S2), as shown in Table 2, the advantage of *Bi-CVS* becomes more evident under heterogeneous variable collections and factor levels. For the continuous and survival endpoints, *Bi-CVS* achieves perfect Level-1 recovery for all p with F_1 scores around 0.95 and very low FPR and FNR. It also consistently outperforms *SepLasso* for the binary endpoint in terms of Level-1 recovery. As p increases, *SepLasso* becomes much less stable: at $p = 120$, its FNR exceeds 0.85 for all three endpoints, whereas *Bi-CVS* maintains F_1 scores of 0.9525, 0.8783, and 0.9553 for the continuous, binary, and survival endpoints, respectively. *SepLasso* is substantially faster computationally, while the run time of *Bi-CVS* increases with the number of candidate variables. Thus, the improved selection accuracy and stability of *Bi-CVS* are obtained at the cost of additional computation. These results demonstrate that *Bi-CVS* remains effective under the heterogeneous study structures that motivate the proposed method, including the setting with $p > n_i$.

5 | Real Clinical Application

Luspatercept is a first-in-class erythroid maturation agent for treating anemia in patients with beta-thalassemia and myelodysplastic syndromes (MDS). In August 2023, the U.S. Food and Drug Administration (FDA) approved Reblozyl (luspatercept-aamt) for the treatment of anemia without previous erythropoiesis-stimulating agent use (ESA-naïve) in adult patients with very low-risk to intermediate-risk MDS who may require regular red blood cell (RBC) transfusions. In this section, we evaluate the performance of the proposed *Bi-CVS* in selecting

prognostic variables using data from four historical clinical trials of luspatercept in patients with MDS. In these historical studies, prognostic variables for covariate adjustment were pre-specified in the SAP by the clinical team. When designing a new luspatercept MDS study, we aim to statistically identify prognostic variables from historical data to support this pre-specification process.

In these four historical studies of luspatercept, summarized in Table 3, MDS-001 and MDS-002 are global, two-arm studies focusing on transfusion-dependent (TD) subjects. MDS-001 is designed for second-line treatment in ESA-refractory patients, whereas MDS-002 targets first-line treatment in ESA-naïve patients. On the other hand, MDS-003 and MDS-004 are single-arm studies conducted only in Asia. MDS-003 enrolls non-transfusion dependent (NTD) patients, while MDS-004 targets TD patients. For each study, the endpoint was defined as a binary response indicator. Since no reliable prior information was available for assigning different weights to the four studies, we used equal weights $w_i = 1/4$ in this case study.

These studies highlight challenges encountered in real-world applications due to their different designs (inclusion/exclusion eligibility) and patient populations, leading to variations in the collection of baseline covariates across studies. For instance, the *RACE* variable is omitted in MDS-003 and MDS-004, which were conducted exclusively in Asia. Similarly, the *RBC* variable is absent in MDS-003 since it enrolls only non-transfusion dependent (NTD) patients. These discrepancies highlight the necessity of the proposed *Bi-CVS* method, which accommodates data heterogeneity and leverages common information across similar studies to identify prognostic variables.

As only a limited number (i.e., 2–4) of stratification factors is considered in clinical trial designs, for ease of use and interpretation, the proposed method allows clinical teams to explicitly specify the number of prognostic variables to be selected. If the optimal number is unknown, it can be determined by tuning the Level-1 constraint parameter η . In studies of luspatercept for treating anemia in MDS patients, the proposed method selected *RBC*, *EPO*, and *IPSSR* to meet the request for three prognostic variables. These variables were also aligned with the factors recommended by the clinical team in the program. The interaction plots between treatment and these three variables in the two-arm trials MDS-001 and MDS-002 are displayed in Figure 3. MDS-003 and MDS-004 did not appear in Figure 3 because, as single-arm studies, they do not permit treatment-by-subgroup comparisons. However, both studies were included in the proposed *Bi-CVS* analysis and contributed to the cross-study prognostic variable selection. The plots show nearly parallel lines across subgroups, suggesting that these variables may primarily act as prognostic rather than treatment-interaction variables in these two randomized studies. This suggests that incorporating these variables in stratified randomization could provide more balanced strata and improve precision for estimating treatment effects.

To formally assess the interactions between treatment and the selected variables, we fitted separate logistic regression models for each variable within each study, including the main effects

TABLE 1 | Simulation results under scenario (S1) with $N = 8$, $n_i = 100$, and $p \in \{20, 90, 120\}$.

Endpoint	Metric	$p = 20$				$p = 90$				$p = 120$			
		Bi-CVS	SepLasso	MTLLasso	MetaLasso	Bi-CVS	SepLasso	MTLLasso	MetaLasso	Bi-CVS	SepLasso	MTLLasso	MetaLasso
Continuous	Level-1 exact	1 (0.00)	0.8700 (0.34)	1 (0.00)	1 (0.00)	1 (0.00)	0.9800 (0.14)	1 (0.00)	0.9900 (0.10)	1 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
	Level-1 recall	1 (0.00)	0.9650 (0.09)	1 (0.00)	1 (0.00)	1 (0.00)	0.9950 (0.04)	1 (0.00)	0.9975 (0.03)	1 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
Binary	FNR	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0.0003 (0.00)	0 (0.00)	0.0019 (0.02)	0 (0.00)	0.0008 (0.01)	0 (0.00)	0 (0.00)
	FPR	0.0206 (0.01)	0.3086 (0.07)	0.0378 (0.02)	0.0231 (0.01)	0.0045 (0.00)	0.1223 (0.03)	0.0074 (0.00)	0.0046 (0.00)	0.0037 (0.00)	0.0957 (0.02)	0.0058 (0.00)	0.0038 (0.00)
	F_1	0.9499 (0.03)	0.5731 (0.06)	0.9119 (0.05)	0.9438 (0.04)	0.9439 (0.03)	0.3944 (0.05)	0.9112 (0.04)	0.9420 (0.03)	0.9385 (0.04)	0.3794 (0.06)	0.9051 (0.04)	0.9364 (0.04)
	Run time (s)	1.9541 (0.22)	0.0714 (0.01)	0.2206 (0.02)	0.7610 (0.04)	30.6824 (2.72)	0.1450 (0.01)	0.4047 (0.04)	1.0467 (0.08)	47.1353 (7.09)	0.1369 (0.01)	0.4430 (0.03)	1.1892 (0.07)
	Level-1 exact	0.9800 (0.14)	0.7400 (0.44)	0.9800 (0.14)	0.1000 (0.30)	0.9600 (0.20)	0.6000 (0.49)	0.9200 (0.27)	0.7400 (0.44)	0.9600 (0.20)	0.6000 (0.49)	0.9800 (0.14)	0.7200 (0.45)
	Level-1 recall	0.9950 (0.04)	0.9325 (0.12)	0.9950 (0.04)	0.5825 (0.24)	0.9900 (0.05)	0.8850 (0.15)	0.9800 (0.07)	0.9200 (0.16)	0.9900 (0.05)	0.9000 (0.12)	0.9950 (0.04)	0.9250 (0.13)
	FNR	0.0374 (0.05)	0.1538 (0.09)	0.0040 (0.03)	0.3697 (0.22)	0.0527 (0.06)	0.3404 (0.12)	0.0157 (0.05)	0.0845 (0.16)	0.0496 (0.07)	0.3981 (0.13)	0.0033 (0.02)	0.0765 (0.12)
	FPR	0.0204 (0.01)	0.2277 (0.07)	0.0380 (0.02)	0.0128 (0.01)	0.0044 (0.00)	0.0751 (0.02)	0.0081 (0.00)	0.0050 (0.00)	0.0030 (0.00)	0.0554 (0.02)	0.0055 (0.00)	0.0052 (0.01)
	F_1	0.9312 (0.04)	0.5747 (0.06)	0.9102 (0.05)	0.6909 (0.20)	0.9168 (0.06)	0.3681 (0.05)	0.8932 (0.08)	0.8869 (0.12)	0.9238 (0.06)	0.3475 (0.06)	0.9090 (0.06)	0.8771 (0.10)
	Run time (s)	2.0452 (0.27)	0.0926 (0.02)	0.2617 (0.03)	4.3175 (1.64)	52.8830 (4.16)	0.1506 (0.01)	0.4592 (0.03)	7.9029 (0.77)	65.3795 (4.18)	0.1500 (0.01)	0.5230 (0.04)	12.4373 (2.30)
Survival	Level-1 exact	1 (0.00)	0.9500 (0.22)	—	0.9400 (0.24)	1 (0.00)	0.9600 (0.20)	—	0.6875 (0.47)	1 (0.00)	0.9800 (0.14)	—	0.1800 (0.39)
	Level-1 recall	1 (0.00)	0.9875 (0.05)	—	0.9850 (0.06)	1 (0.00)	0.9900 (0.05)	—	0.8698 (0.22)	1 (0.00)	0.9950 (0.04)	—	0.5550 (0.32)
	FNR	0 (0.00)	0.0073 (0.02)	—	0.0121 (0.05)	0.0016 (0.01)	0.0313 (0.04)	—	0.1270 (0.22)	0.0007 (0.01)	0.0367 (0.05)	—	0.4563 (0.34)
	FPR	0.0202 (0.01)	0.2545 (0.05)	—	0.0280 (0.01)	0.0038 (0.00)	0.0809 (0.02)	—	0.0030 (0.00)	0.0024 (0.00)	0.0608 (0.01)	—	0.0016 (0.01)
	F_1	0.9515 (0.03)	0.6151 (0.05)	—	0.9276 (0.04)	0.9519 (0.03)	0.4865 (0.05)	—	0.8767 (0.15)	0.9603 (0.02)	0.4843 (0.05)	—	0.6118 (0.33)
	Run time (s)	1.9556 (0.44)	0.1188 (0.01)	—	4.9593 (0.55)	52.4157 (2.20)	7.7822 (0.62)	—	49.1526 (11.88)	73.0682 (3.65)	4.4025 (0.16)	—	57.5427 (16.17)

Note: Values are reported as mean (SD) over 100 simulation replications. Run time is reported in seconds. In the binary setting with $p = 20$, *MetaLasso* frequently selected fewer than the target number of four variables, resulting in lower Level-1 recovery. Three replications were excluded because the final model fit failed.

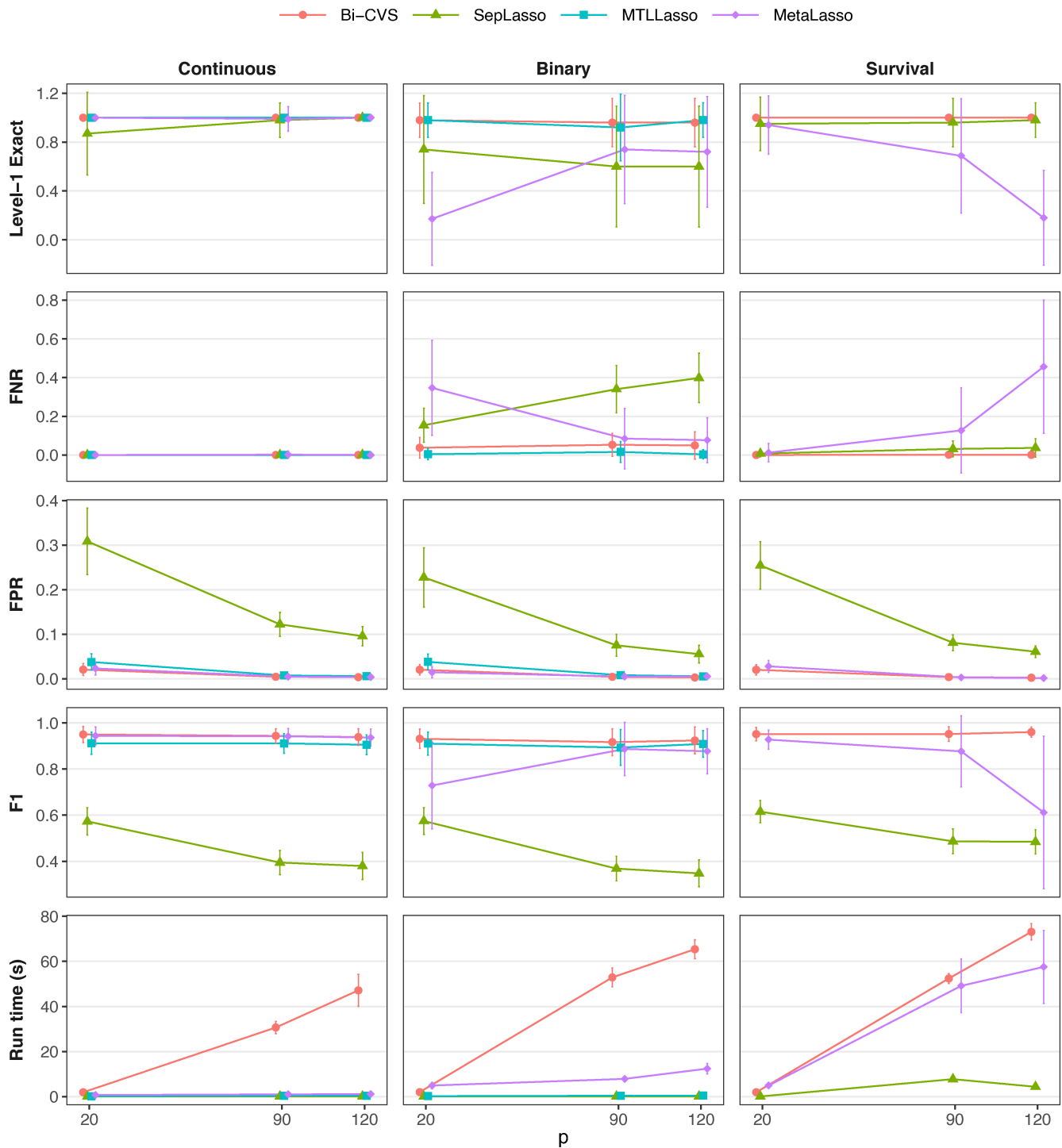


FIGURE 2 | Comparison of variable selection performance and computation time under scenario (S1). Points represent mean values and error bars represent one standard deviation across 100 simulation replications. Columns correspond to continuous, binary, and survival endpoints, and rows correspond to Level-1 exact recovery, FNR, FPR, F_1 , and run time, respectively.

of treatment and the selected variable and their interaction. The interaction effect is assessed using a likelihood ratio test comparing the interaction model with the corresponding main-effects model. The results are reported in Table 4. From these results, one can see that none of the treatment-by-variable interactions is statistically significant in either study (all p -values ≥ 0.2254). Thus, these results are consistent with the approximately parallel lines in Figure 3 and suggest that the selected variables are prognostic.

6 | Discussion

This work was motivated by a practical problem in the design of new clinical trials, where regulatory guidance recommends pre-specifying key prognostic variables in the statistical analysis plan. In collaboration with Bristol Myers Squibb, we considered the setting in which a new trial is informed by multiple historical studies that may differ in study design, patient population, and

TABLE 2 | Simulation results under scenario (S2) with $N = 8$, $n_i = 100$, and $p \in \{20, 90, 120\}$.

Endpoint	Metric	$p = 20$		$p = 90$		$p = 120$	
		Bi-CVS	SepLasso	Bi-CVS	SepLasso	Bi-CVS	SepLasso
Continuous	Level-1 exact	1 (0.00)	0.7200 (0.45)	1 (0.00)	0.9000 (0.30)	1 (0.00)	0.9600 (0.20)
	Level-1 recall	1 (0.00)	0.9300 (0.11)	1 (0.00)	0.9725 (0.09)	1 (0.00)	0.9900 (0.05)
	FNR	0 (0.00)	0 (0.00)	0 (0.00)	0.0172 (0.06)	0 (0.00)	0.8574 (0.09)
	FPR	0.0169 (0.01)	0.3467 (0.07)	0.0042 (0.00)	0.0050 (0.00)	0.0028 (0.00)	0.1442 (0.03)
	F_1	0.9614 (0.03)	0.5518 (0.04)	0.9495 (0.03)	0.9326 (0.07)	0.9525 (0.03)	0.0622 (0.04)
	Run time (s)	1.8647 (0.17)	0.1409 (0.01)	28.8701 (2.65)	0.5090 (0.03)	40.3947 (3.56)	0.4621 (0.03)
Binary	Level-1 exact	0.9800 (0.14)	0.6800 (0.47)	0.6800 (0.47)	0.3200 (0.47)	0.7800 (0.42)	0.3600 (0.48)
	Level-1 recall	0.9950 (0.04)	0.9150 (0.13)	0.9000 (0.18)	0.7800 (0.20)	0.9450 (0.10)	0.8050 (0.18)
	FNR	0.0670 (0.08)	0.1796 (0.10)	0.1209 (0.13)	0.2430 (0.18)	0.1094 (0.11)	0.9165 (0.08)
	FPR	0.0121 (0.01)	0.2292 (0.06)	0.0049 (0.00)	0.0070 (0.00)	0.0039 (0.00)	0.0860 (0.03)
	F_1	0.9359 (0.05)	0.5694 (0.05)	0.8560 (0.16)	0.7561 (0.18)	0.8783 (0.10)	0.0505 (0.04)
	Run time (s)	2.6126 (0.53)	0.5386 (0.13)	34.8036 (6.53)	0.7175 (0.11)	53.2139 (3.34)	0.6782 (0.04)
Survival	Level-1 exact	1 (0.00)	0.8400 (0.37)	1 (0.00)	0.9800 (0.14)	1 (0.00)	0.9200 (0.27)
	Level-1 recall	1 (0.00)	0.9600 (0.09)	1 (0.00)	0.9950 (0.04)	1 (0.00)	0.9800 (0.07)
	FNR	0.0018 (0.01)	0.0091 (0.02)	0.0009 (0.01)	0.0038 (0.03)	0.0010 (0.01)	0.9052 (0.07)
	FPR	0.0179 (0.01)	0.2867 (0.08)	0.0034 (0.00)	0.0036 (0.00)	0.0027 (0.00)	0.1035 (0.02)
	F_1	0.9586 (0.02)	0.6014 (0.06)	0.9588 (0.03)	0.9551 (0.04)	0.9553 (0.03)	0.0529 (0.04)
	Run time (s)	1.8464 (0.20)	0.1687 (0.02)	49.7179 (3.55)	0.7068 (0.03)	60.8319 (3.78)	0.6802 (0.04)

Note: Values are reported as mean (SD) over 100 simulation replications. Run time is reported in seconds.

collected baseline covariates. To address this challenge, we developed the *Bi-CVS* method within a multi-task learning framework to identify a prespecified number of prognostic variables across heterogeneous historical studies.

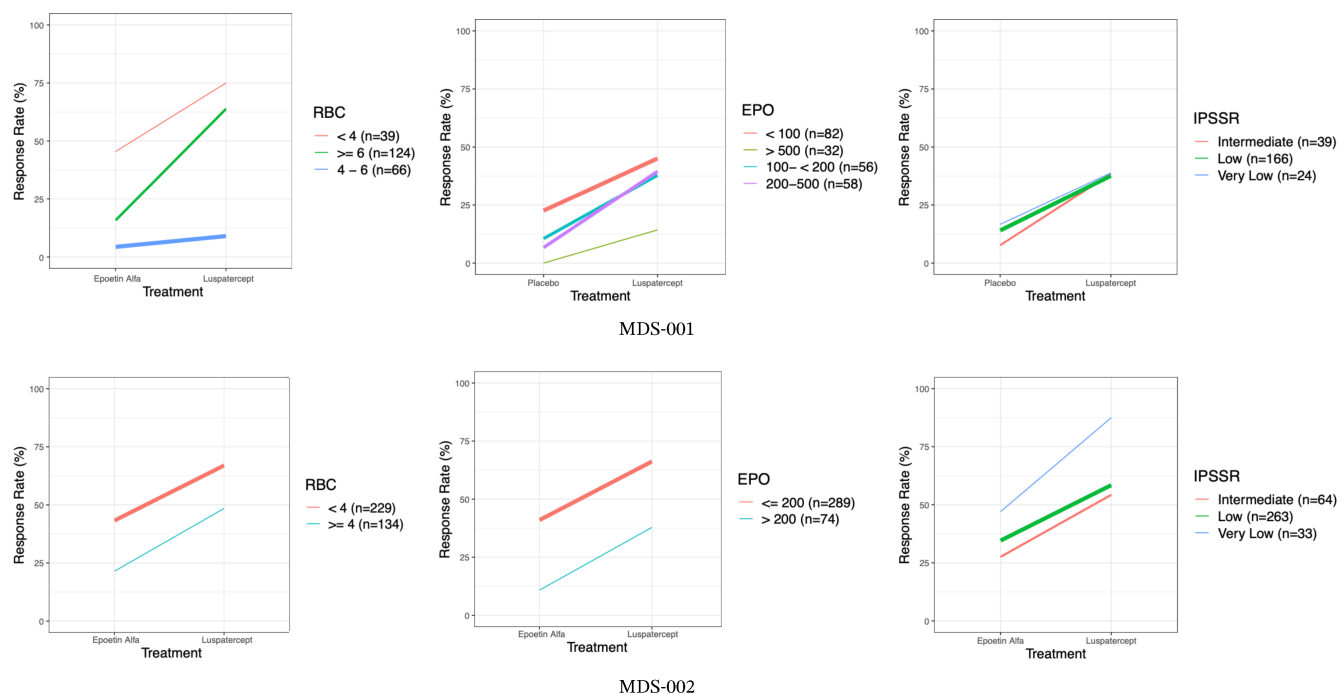
The proposed method is designed for settings in which historical studies are related but not identical. It accommodates partially overlapping covariate sets across studies and allows common categorical variables to have different observed levels in different studies. In addition, study-specific weights and within-study penalization allow the selected prognostic variables to affect the outcome differently across studies. To implement this framework, we developed a backward selection algorithm that directly

returns a user-specified number of prognostic variables, making it easier to interpret and apply. The simulation studies and real-data application further support the effectiveness of the proposed approach.

There are several promising directions for future research. First, it would be valuable to extend the framework to settings with a large number of potentially relevant historical studies. Although incorporating more studies may improve information borrowing, it also increases computational burden and may reduce performance when some studies are only weakly related to the target study. One possible direction is to develop study-selection or study-weighting strategies based on similarity to the new study.

TABLE 3 | Summaries of luspatercept MDS studies.

Study	Design	Subject	Population	Treatment	Region	Variables
MDS-001	Two-arm	ESA-refractory, RS+, TD	229	Luspatercept (153) or Placebo (76)	Global	AGE, GENDER, RACE, RBC, EPO, MDS, T2DIAG, ECOG, IPSSR, PLATE
MDS-002	Two-arm	ESA-naïve, RS+ & RS-, TD	363	Luspatercept (182) or Epoetin Alfa (181)	Global	AGE, GENDER, RACE, RBC, RS, EPO, MDS, T2DIAG, ECOG, IPSSR, PLATE
MDS-003	Single-arm	RS+ & RS-, NTD	21	Luspatercept	Japan	AGE, GENDER, RS, EPO, MDS, T2DIAG, ECOG, IPSSR, PLATE, ESA
MDS-004	Single-arm	RS+, TD	30	Luspatercept	China and Japan	AGE, GENDER, RBC, EPO, MDS, T2DIAG, ECOG, IPSSR, PLATE, ESA

**FIGURE 3** | Interaction plots displaying the average response rates for placebo versus luspatercept in MDS-001 and for epoetin alfa versus luspatercept in MDS-002, segmented by patient subgroups for three selected prognostic variables: RBC, EPO, and IPSSR. The absence of MDS-003 and MDS-004 from these plots is due to their nature as single-arm trials.**TABLE 4** | Likelihood ratio tests for interactions between treatment and the selected prognostic variables.

Variable	MDS-001		MDS-002	
	χ^2	<i>p</i>	χ^2	<i>p</i>
RBC	0.2046	0.6511	0.1341	0.7142
EPO	1.4695	0.2254	0.6858	0.4076
IPSSR	0.4924	0.4829	0.0220	0.8820

Another is to formulate the incorporation of additional studies as a data-adaptive or active learning problem. Second, the theoretical analysis is currently restricted to fixed p , while the computational burden increases as p grows. More scalable algorithms and corresponding theoretical developments are needed for higher-dimensional settings. Third, the proposed framework may be extended to accommodate other data structures, such as longitudinal data, as well as applications in genomics or finance.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The code and simulated datasets used in the simulation studies are provided as [Supporting Information](#) with this article. The real-world clinical trial data analyzed in the application are owned by Bristol Myers Squibb (BMS) and cannot be shared publicly by the authors due to data use restrictions and patient privacy requirements. Requests for access should be directed to BMS through its applicable data-sharing process.

References

1. US Food and Drug Administration, "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products," 2023.

2. G. M. Raab, S. Day, and J. Sales, “How to Select Covariates to Include in the Analysis of a Clinical Trial,” *Controlled Clinical Trials* 21, no. 4 (2000): 330–342.
3. R. Caruana, “Multitask Learning,” *Machine Learning* 28 (1997): 41–75.
4. Y. Zhang and Q. Yang, “An Overview of Multi-Task Learning,” *National Science Review* 5, no. 1 (2018): 30–43.
5. A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-Task Feature Learning,” in *Advances in Neural Information Processing Systems* (MIT Press, 2006), 19.
6. G. Obozinski, B. Taskar, and M. Jordan, “Multi-Task Feature Selection,” Statistics Department, UC Berkeley, Tech. Rep., 2006, 2 (2.2), 2.
7. G. Swirszcz and A. C. Lozano, “Multi-Level Lasso for Sparse Multi-Task Regression,” in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (Omnipress, 2012), 361–368.
8. Q. Li, S. Wang, C. C. Huang, M. Yu, and J. Shao, “Meta-Analysis Based Variable Selection for Gene Expression Data,” *Biometrics* 70, no. 4 (2014): 872–880.
9. S. J. Pocock and R. Simon, “Sequential Treatment Assignment With Balancing for Prognostic Factors in the Controlled Clinical Trial,” *Biometrics* 31 (1975): 103–115.
10. C. J. Weir and K. R. Lees, “Comparison of Stratification and Adaptive Methods for Treatment Allocation in an Acute Stroke Clinical Trial,” *Statistics in Medicine* 22, no. 5 (2003): 705–726.
11. Committee for Medicinal Products for Human Use (CHMP), “Guideline on Adjustment for Baseline Covariates in Clinical Trials,” 2015.
12. R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 58, no. 1 (1996): 267–288.
13. M. Yuan and Y. Lin, “Model Selection and Estimation in Regression With Grouped Variables,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 68, no. 1 (2006): 49–67.
14. A. Jalali, S. Sanghavi, C. Ruan, and P. Ravikumar, “A Dirty Model for Multi-Task Learning,” in *Advances in Neural Information Processing Systems* (MIT Press, 2010), 23.
15. M. Stone, “Cross-Validatory Choice and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 36, no. 2 (1974): 111–133.
16. D. Lin, D. P. Foster, and L. H. Ungar, *A Risk Ratio Comparison of ℓ_0 and ℓ_1 Penalized Regressions. Technical Report* (University of Pennsylvania, 2010).
17. W. Xie and X. Deng, “Scalable Algorithms for the Sparse Ridge Regression,” *SIAM Journal on Optimization* 30, no. 4 (2020): 3359–3386.
18. B. K. Natarajan, “Sparse Approximate Solutions to Linear Systems,” *SIAM Journal on Computing* 24, no. 2 (1995): 227–234.
19. T. Zhang, “Adaptive Forward-Backward Greedy Algorithm for Sparse Learning With Linear Models,” in *Advances in Neural Information Processing Systems* (MIT Press, 2008), 21.
20. T. Zhang, “On the Consistency of Feature Selection Using Greedy Least Squares Regression,” *Journal of Machine Learning Research* 10, no. 3 (2009): 555–568.
21. T. M. Therneau, “How Many Stratification Factors Are “Too Many” to Use in a Randomization Plan?,” *Controlled Clinical Trials* 14, no. 2 (1993): 98–108.
22. A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics* 12, no. 1 (1970): 55–67.
23. H. Zou and T. Hastie, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 67, no. 2 (2005): 301–320.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supplementary Figure 1:** Overview of R Shiny app. **Data S1:** Supporting Information.