



Integrative Analysis of Site-Specific Parameters with Nuisance Parameters on the Common Support

Ho Yi Alexis Ho¹ · Shuoxun Xu¹ · Xinzhou Guo¹

Received: 30 September 2023 / Revised: 2 February 2024 / Accepted: 17 March 2024
© The Author(s) 2024

Abstract

High-throughput technologies in bioscience have pushed us into an era with high dimensionality. Swamped by thousands of predictors, detecting the valuable signal from the noise in clinical studies becomes challenging. As a common strategy, integrative analysis utilizing similarities across multiple studies might help lift the curse of dimensionality and enhance statistical power. However, due to the growing concern about individual data privacy, data-sharing constraints are often imposed in integrative analysis. These might lead to results inequivalent to ones without sharing constraints and reduce statistical power in integrative analyses. In this paper, built on Abess, we propose an integrative analysis method to estimate the site-specific parameters in the presence of high dimensional nuisance parameters in multi-site studies. Implemented with a carefully designed $L_{2,0}$ penalization on nuisance parameters, the proposed method satisfies both the DataSHIELD constraint, which only allows the transmission of summary statistics from sites, and the equivalence property that the solution is exactly the same as the solution merging all datasets into one on a single location. Assuming the nuisance parameters share a common support, the proposed method has support recovery and selection consistency with high probability and exhibits improved estimation accuracy on the site-specific parameters and low computational cost in numerical experiments. We demonstrate the merit of the proposed method by investigating the relationship between the CD8 T cell count and the treatment effect of zidovudine-incorporated therapy in the AIDS Clinical Trials Group Study 175.

Keywords Computational cost · DataSHIELD · Equivalence · $L_{2,0}$ penalization on nuisance parameters · Multi-site · High dimensions

Extended author information available on the last page of the article

Published online: 29 April 2024

1 Introduction

1.1 Integration Analysis Under DataSHIELD Constraint

Advancements in high-throughput technologies enable bioscientists to access data from numerous angles within one biological sample, and hundreds, even thousands of variables associated with the genome, transcriptome, proteome, metabolome, epigenome, etc., are of interest [1]. Extracting and statistically analyzing informative features are a primary step in understanding the biological mechanisms of diseases, such as HIV [2–4]. However, because of the limited observational sample size, detecting the signal of informative features is rather challenging due to the high dimensionality [5]. In practice, a natural solution to address the curse of dimensionality is to increase the sample size by merging datasets from different sources, often known as integrative analysis in the literature [5, 6]. Such integration is regular in evidence-based medicine apart from the group-centered studies [7], as valuable medical topics are often repetitively examined by more than one research unit and share similarities across different locations [8]. Still, the discrepancy exists in the patient population, and heterogeneity between various studies becomes a significant challenge in integrative analysis [5, 9]. Another obstacle in the integrative analysis is related to data-sharing constraints; analysts might not be able to share individual-level data because of legal and privacy concerns. For instance, the patient-level medical information linked with electronic health records (EHR) usually cannot go past the firewall of its associated hospital [5].

Facing such data-sharing constraints, [10] proposed a widely adopted individual privacy-preserving framework, DataSHIELD, for integrative analysis. The mechanism of DataSHIELD is to pass only summary statistics from decentralized local nodes to the central node in integrative analysis. However, current DataSHIELD-supported approaches under the high-dimensional setting often fail to consider or cannot be easily extended to accommodate cross-site heterogeneity without sacrificing statistical efficiency; see [5, 6]. These examples include, but are not limited to, the aggregated debiased lasso estimator adopted by [11–13] where local debiasing might incur additional estimation errors. Other works, such as [14, 15], can avoid efficiency loss but only work for cross-site homogeneous scenarios and demand successive communications between local machines and central nodes, which could waste time and resources.

Recently, several improvements have been made to accommodate cross-site heterogeneity in integrative analysis within the DataSHIELD framework [5, 6], but these require the calculation of the estimator at local sites and an approximation of the loss function. Such an approximation can lead to inequivalent solutions in contrast with the solutions obtained by merging all datasets into one, followed by potentially associated accuracy and efficiency loss. In addition, the DSILT Algorithm proposed by [5] requires the summary statistics to be updated and transmitted frequently, which may pose further privacy concerns besides increased computational complexity. Relevant work has also been considered

under the names of distributed learning [16] and federated learning [17–20]. However, similar to the existing literature in integrative analysis, they often fail to address one or some of the issues concerning high dimensionality, site heterogeneity, data privacy, computational cost, estimation accuracy, and equivalence, and we aim to bridge the gap.

In this paper, we focus on the DataSHIELD constraint and propose a $L_{2,0}$ method to conduct integrative analysis. The proposed method can naturally accommodate high dimensionality and site heterogeneity with low computational cost, and improve the estimation accuracy by utilizing the common supports across different sites and implementing $L_{2,0}$ penalization solely on nuisance parameters. The most distinctive feature of our algorithm is that it can achieve equivalence (in contrast with approximated ones) under the DataShield constraint, which means the results produced by our algorithm are exactly the same as the ones if all datasets across different sites have been merged into one. The equivalence property helps eliminate the concern of potential estimation accuracy and efficiency loss induced by approximation algorithms for integrative analysis under the data-sharing constraint. We introduce the model setting and outline of the proposed method, particularly the common support assumption and the $L_{2,0}$ penalization in Sects. 1.2 and 1.3, respectively.

Notations: $|\cdot|$ denotes the size of the set. $[i]$ represents the set $\{1, \dots, i\}$. Suppose a, b are constants. $a[i] := \{a, \dots, ai\}$ and $[i] - b := \{1 - b, \dots, i - b\}$. That is, we operate on each element in the set. $S_1 \setminus S_2$ symbolizes the set difference operation. We define $S_1 \times S_2$ as the cartesian products of two sets S_1 and S_2 . $X_{[S_1 \times S_2]}$ means the submatrix whose entires are in the k th row and ℓ th column of matrix X , where $k \in S_1$ and $\ell \in S_2$. Inspired by [21], we modified their notations and practiced as follows. Define selected set $\mathcal{A} = \{j \in [p] : \|\beta_{G_j}\|_2 \neq 0\}$. The unselected set is $\mathcal{I} = [p] \setminus \mathcal{A} = \mathcal{A}^c$. We let $\beta_{\mathcal{A}} = (\beta_{G_j}, j \in \mathcal{A}) \in \mathcal{R}^{M \cdot |\mathcal{A}|}$. That is, the dimension of $\beta_{\mathcal{A}}$ is the scalar product between the total number of sites M and the set size $|\mathcal{A}|$. We define $X_{\mathcal{A}} = (X_{G_j}, j \in \mathcal{A}) \in \mathbb{R}^{n \times (M \cdot |\mathcal{A}|)}$ and denote β^* as the true regression coefficients. The true subsets of groups is $\mathcal{A}^* = \{j \in [p] : \|\beta_{G_j}^*\|_2 \neq 0\}$. $\mathcal{I}^* = (\mathcal{A}^*)^c$.

1.2 Model Statement

We consider an integrative analysis problem with multiple datasets of linear models, a setting broadly considered in practice; see [22, 23]. Suppose there are M independent studies. The m -th study contains n_m random observations on the outcomes $y^{(m)} = (y_1^{(m)}, \dots, y_{n_m}^{(m)})^\top$, vector $D^{(m)} = (D_1^{(m)}, \dots, D_{n_m}^{(m)})^\top \in \mathbb{R}^{n_m}$ and covariate matrix $X^{(m)} = (X_1^{(m)\top}, \dots, X_{n_m}^{(m)\top})^\top \in \mathbb{R}^{n_m \times p}$. Within each study, we assume the same regression model:

$$y_i^{(m)} = D_i^{(m)} \cdot \alpha^{(m)} + X_i^{(m)} \beta^{(m)} + \epsilon_i^{(m)}, \quad i = 1, \dots, n_m \tag{1}$$

where $\alpha^{(m)} \in \mathcal{R}$ denotes the site-specific parameter of our interest. Here, we focus on a one-dimensional site-specific parameter for the sake of presentation, but the model can be naturally extended to multi-dimensional site-specific parameters. The

coefficient vector $\beta^{(m)} = (\beta_1^{(m)}, \dots, \beta_p^{(m)})^\top$ represents nuisance parameters, which are out of our interest but used for model adjustment in practice. $\epsilon_i^{(m)}$ are i.i.d. error terms satisfying $\mathbb{E}(\epsilon_i^{(m)} | D_i^{(m)}, X_i^{(m)}) = 0$.

One motivating example of Model (1) is the estimation of treatment effects in multi-site observational studies where the potential confounder bias needs to be accounted for by appropriate statistical methods, such as matching methods [24, 25] and AIPWE [26]. Here, as suggested by [27, 28], we consider using a regression model to adjust the potential confounder bias and let $D^{(m)}$ denote the treatment indicator and $X^{(m)}$ denote the potential confounder of the m -th site. Then, the site-specific parameter $\alpha^{(m)}$ is the treatment effect of the m -th site, which is of our interest, and $\beta^{(m)}$ is the confounder effect, a nuisance parameter. Other motivating examples include but are not limited to the repeated measurement design, where a subject might be measured multiple times across different sites [29], and the multiple-measurement-vector (MMV) problem, where signals are collected from different sources [30].

Based on Model (1), to borrow similarities across different sites to improve estimation accuracy in integrative analysis, we consider a common support assumption that the sparsity sets are the same across different sites. Sparsity is a widely adopted assumption to ensure the identifiability of the model in high dimensions [31]; i.e., $p > n_m$. Although the parameters might not be the same across different sites due to site heterogeneity, the sparsity of the nuisance parameter might be the same in many practical applications. Take the multi-site observational study as an example; though the specific effects of the confounders might be different, the true confounders (features) are often the same due to the similarity of patients' preferences in choosing the drugs regardless of the sites. Other examples naturally bearing the common support assumption include the repeated measurement design and the MMV problem where for each subject or signal, measurement across different sites are expected to have the same sparsity set [32–34]. In specific, let

$$\text{supp}(\beta^{(m)}) = \{j : \beta_j^{(m)} \neq 0\}. \tag{2}$$

denote the active set. The common support assumption means that

$$\text{supp}(\beta^{(1)}) = \dots = \text{supp}(\beta^{(m)}) = \dots = \text{supp}(\beta^{(M)}); \tag{3}$$

$$s = |\text{supp}(\beta^{(m)})| < n. \tag{4}$$

Note that here, we impose the common support assumption solely on the nuisance parameter. This is because the site-specific parameter is what we wish to study and to improve the estimation accuracy of the parameter of interest, site-specific parameters should be retained in the reduced model rather than being screened out during the feature selection [35].

The main difference between the common support assumption considered in this paper and the similar parameter assumption considered in the fusion method [11, 36] is where the sparsity arises. Specifically, the common support assumption is

w.r.t. the sparsity of the parameters, while the similar parameter assumption is w.r.t. the sparsity of the distance between the parameters, which is similar to the difference between lasso and fused lasso [37]. Both the sparsity of the parameters and the sparsity of the distance between the parameters have practical implications, and we might adopt one or both in real-world scenarios; e.g., the sparsity in genomics [38] and the sparsity in the distance between parameters in time-varying/spatial data [39]. In this paper, we focus on the common support assumption, which induces the same sparsity of the parameter across different sites.

1.3 Outline of the Proposed Method

Besides passing only summary statistics from the local sites, our proposed method consists of two key elements. The first key element of the proposed method is that we apply the $L_{2,0}$ penalization to induce the same support set over the nuisance parameter. The second key element of the proposed method is that we place the $L_{2,0}$ penalization solely on the nuisance parameter but not on the site-specific parameter.

For the first key element, we consider $L_{2,0}$ instead of $L_{2,1}$, another widely adopted strategy to induce the same support, for the following two reasons. First, it is well known that $L_{2,1}$ suffers from the selection bias and over-shrinkage of significant coefficients, while $L_{2,0}$ is more favorable as it allows an explicit presentation of support size [21, 40]. Second, based on $L_{2,0}$, with a carefully designed algorithm to assemble the summary statistics, we can address inequivalence issues in those approximation integrative analysis methods based on $L_{2,1}$ under the DataSHIELD constraint. The second key element, penalization solely on the nuisance parameter, is due to the following two reasons. The first reason is such partial penalization can help improve model interpretation. Take the above multi-site observational study as an example; it is hard for researchers to conclude or explain the treatment efficacy if the treatment indicator is excluded from their selected model [41]. The second reason is that keeping the parameter of interest in the model can help improve estimation accuracy. As argued by [35], prior information should be taken into account in feature selection. In particular, the parameters deemed important, such as the parameter of interest in this paper, should be retained in the selected model to improve estimation accuracy.

The rest of the paper will be organized as follows. We will first review the single-site algorithm proposed by [21] and then present ours, enabling multi-site functioning with penalization solely on the nuisance parameter in Sect. 2. The theoretical properties will be in Sect. 3, followed by the simulation and the real-data application in Sects. 4 and 5, respectively. At the end, we will have a summary in Sect 6.

2 Methodology

2.1 Innovative Use of Splicing Approach

As discussed in Sect. 1.3, we aim to apply $L_{2,0}$ to induce the same support across different sites. It is well known that $L_{2,0}$ leads to an intractable nonconvex problem, and

among the few existing computational methods for $L_{2,0}$, we consider the single-site best subset of groups selection (BSGS) algorithm - Abess, which adopts a splicing approach proposed by [21]. In contrast to traditional feature selection methods, which focus on each variable individually, BSGS incorporates the grouping information of variables and selects features at the group level. By employing the splicing technique, the algorithm iteratively includes the significant groups and discards the nonessential ones, enhancing the interpretability of the outcome variable. Under mild assumptions, the algorithm has been proven to possess polynomial complexity with a high probability of determining the optimal subsets of groups, even in high-dimensional feature spaces.

In specific, [21] assumed a linear model composed of J non-overlapping groups, referred to as a group linear model. The model is formulated as follows:

$$y = \sum_{j=1}^J X_{G_j} \beta_{G_j} + \varepsilon. \tag{5}$$

They denote $y \in \mathbb{R}^n$ as the outcome variable and represent the j th group’s regressor matrix as $X_{G_j} \in \mathbb{R}^{n \times p_j}$, where p_j is size of the j th group. They define $\beta_{G_j} \in \mathbb{R}^{p_j}$ as the j th group’s regression coefficients and $\varepsilon \in \mathbb{R}^n$ as the random error term. G_j refers to a collection of indices associated with predictors that belong to the j th group. Additionally, $\cup_{j=1}^J G_j = [p]$, where G_i and G_j are disjoint for all $i \neq j$. In the simplest case, the model converts to the ordinary linear model when the group size $p_j = 1$ for all $j \in [J]$.

To use splicing to address the problem of our interest, we need to categorize the same variable across different sites as a group in an appropriate way. Suppose we are indifferent between the site-specific and nuisance parameters. We can modify $y = (y^{(1)\top}, \dots, y^{(M)\top})^\top \in \mathbb{R}^n$, $\beta_{G_j} = (\beta_j^{(1)}, \dots, \beta_j^{(M)})^\top \in \mathbb{R}^M$, and $\alpha = (\alpha^{(1)}, \dots, \alpha^{(M)})^\top \in \mathbb{R}^M$. However, because of the unmatched dimensionality, one will fail to construct the proper setting when assigning X_{G_j} as a general matrix previously permitted by [21]. Such a challenge motivates us to consider a special regressor matrix with a diagonal block structure. Formally, we design $X_{G_j} = \text{diag}(X_j^{(1)}, \dots, X_j^{(M)}) \in \mathbb{R}^{n \times M}$ for $j \in [p]$. $D = \text{diag}(D^{(1)}, \dots, D^{(M)}) \in \mathbb{R}^{n \times M}$, where $D^{(m)} \in \mathbb{R}^{n_m}$, $\alpha^{(m)} \in \mathbb{R}$, $\beta_j^{(m)} \in \mathbb{R}$ for $j \in [p]$.

Equivalently, we can write the group linear model in its general matrix form. The idea is that we append the treatment indicator to each design matrix of biomarkers on each site, intending to show the indifference between the nuisance and site-specific parameters under our temporary assumption.

$$y = \check{X} \check{\beta} + \varepsilon, \tag{6}$$

where $\check{\beta} = (\check{\beta}^{(1)\top}, \dots, \check{\beta}^{(M)\top})^\top \in \mathbb{R}^{M(p+1)}$ and $\check{\beta}^{(m)} = (\beta^{(m)\top}, \alpha^{(m)\top})^\top$ for $m \in [M]$. $\check{X} = \text{diag}(\check{X}^{(1)}, \dots, \check{X}^{(M)}) \in \mathbb{R}^{n \times (M(p+1))}$ and $\check{X}^{(m)} = [X^{(m)}, D^{(m)}] \in \mathbb{R}^{n_m \times (p+1)}$.

Suppose the entire dataset resides at a single site, and we allow for penalizing site-specific parameters. The feature selection and parameter estimation can be performed using the algorithm introduced by [21] after the problem reformulation.

With initially selected predictors, the algorithm repetitively solves an ordinary least square problem with the objective function

$$\min_{\check{\beta} \in \mathbb{R}^{M(p+1)}} \frac{1}{2n} \|y - \check{X}\check{\beta}\|_2^2 \quad \text{s.t. } \|\check{\beta}\|_{2,0} \leq T, \tag{7}$$

followed by a splicing procedure. That is, the algorithm exchanges the predictors from selected and unselected sets to check for further improvement in the loss.

Yet, there still exist challenges to address the problem of our interest even after we modify [21]’s algorithm by Eq. (7). First, the algorithm in [21] does not distinguish between the site-specific and nuisance parameters, and α is penalized while it is of interest and deemed important. This might lead to less interpretable and accurate results. Second, the algorithm in [21] does not consider the situation of interest where data are stored at multiple sites and processed under the DataSHIELD framework. Extending the single-site algorithm in [21] to multiple sites under the DataSHIELD framework is desired but challenging. In specific, issues arise when the algorithm attempts to compute the optimal solutions of (7). For instance, we need to figure out how one of the oracle solutions,

$$\check{\beta}_{\mathcal{A}^k}^k = (\check{X}_{\mathcal{A}^k}^\top \check{X}_{\mathcal{A}^k})^{-1} \check{X}_{\mathcal{A}^k}^\top y,$$

in the k -th iteration will be computed if only the transmission of summary statistics is permitted.

2.2 Proposed Method

Built on the innovative use of splicing in Sect. 2.1, we propose an integrative analysis algorithm to estimate the site-specific parameters. Specifically, we first modify the objective function and derive an updated solution when $L_{2,0}$ penalization is solely on the nuisance parameter. Then, we develop Algorithm 1 to assemble the summary statistics and outputs requested in the solution of the modified $L_{2,0}$ objective function. In the end, based on Algorithm 1, we formally introduced Algorithm 2, which satisfies the DataSHIELD framework in a multi-site setting without penalizing the site-specific parameter.

To start with, we state the updated objective function

$$\min_{\beta \in \mathbb{R}^{Mp}, \alpha \in \mathbb{R}^M} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 \quad \text{s.t. } \|\beta\|_{2,0} \leq T, \tag{8}$$

where $\beta = [\beta^{(1)\top}, \dots, \beta^{(M)\top}]^\top \in \mathbb{R}^{Mp}$; $\alpha = [\alpha^{(1)}, \dots, \alpha^{(M)}]^\top \in \mathbb{R}^M$, $n = \sum_{m=1}^M n_m$, $X = \text{diag}(X^{(1)}, \dots, X^{(M)}) \in \mathbb{R}^{n \times (Mp)}$, $D = \text{diag}(D^{(1)}, \dots, D^{(M)}) \in \mathbb{R}^{n \times M}$.

Namely, we consider a multi-site linear regression with block diagonal design matrices X and D . The site-specific parameter α is not subject to penalization as we no longer treat β_{G_j} and α_{G_j} as indistinguishable. While the presence of any β_{G_j} for all $j \in [p]$ is indeterminate, we always guarantee the inclusion of α in the model, which we refer to as the concept of conditional feature screening. Since the penalization

term depends on β purely, given β , the optimization of α exclusively pertains to the first term, which is differentiable. Solving it gives us the solution of α . That is,

$$\alpha = (D^T D)^{-1} D^T (y - X\beta). \tag{9}$$

In turn, once we know the optimal solution of α , we can solve for the one for β by splicing. We substitute the given α into (8). Some algebra turns the problem to

$$\min_{\beta \in \mathbb{R}^{M^p}} \frac{1}{2n} \|\hat{y} - \hat{X}\beta\|_2^2 \text{ s.t. } \|\beta\|_{2,0} \leq T, \tag{10}$$

$$\text{where } H = (I - D(D^T D)^{-1} D^T), \tag{11}$$

$$\hat{y} = Hy; \hat{X} = HX. \tag{12}$$

Remark Problem (10) is exactly the question proposed in (7). We can directly apply [21]’s results after the variable transformations. Since H is idempotent and symmetric, we will see cancellations in our final solutions.

However, like [21], we still face the obstacle of solving $\beta_{\mathcal{A}^\circ}^\circ$, the optimal value of β under the optimal selected set \mathcal{A}° , without knowing $X_{\mathcal{A}^\circ}$ (the sub-matrix of X under the optimal selected set \mathcal{A}°) explicitly. Hence, we are motivated to design local summary statistics containing sufficient information to restore our interested variables β^k , α^k , and \mathcal{A}^k of the k -th iteration of splicing leading to the optimal solution in the center. Since for any index set \mathcal{A}^k ,

$$\begin{aligned} \beta_{\mathcal{A}^k}^k &= (X_{\mathcal{A}^k}^T H X_{\mathcal{A}^k})^{-1} X_{\mathcal{A}^k}^T H y \\ &= (X_{\mathcal{A}^k}^T X_{\mathcal{A}^k} - X_{\mathcal{A}^k}^T D (D^T D)^{-1} D^T X_{\mathcal{A}^k})^{-1} (X_{\mathcal{A}^k}^T y - X_{\mathcal{A}^k}^T D (D^T D)^{-1} D^T y), \end{aligned} \tag{13}$$

the involved pairwise products,

$$X_{\mathcal{A}^k}^T X_{\mathcal{A}^k}, X_{\mathcal{A}^k}^T y, X_{\mathcal{A}^k}^T D, D^T D, D^T y,$$

can be assembled by their local counterparts, respectively. That is summary statistics

$$X_{\mathcal{A}^k}^{(m)T} X_{\mathcal{A}^k}^{(m)}, X_{\mathcal{A}^k}^{(m)T} y^{(m)}, X_{\mathcal{A}^k}^{(m)T} D^{(m)}, D^{(m)T} D^{(m)}, D^{(m)T} y^{(m)},$$

sent from local sites $m = 1, \dots, M$. Such a procedure, computing distributively and then assembling in the center, is achievable due to the diagonal design of the matrix X in our problem. As an illustrative example, we give the guidelines of constructing $X_{\mathcal{A}^k}^T X_{\mathcal{A}^k}$ from $X_{\mathcal{A}^k}^{(m)T} X_{\mathcal{A}^k}^{(m)}$ in Algorithm (1). For each site m , we generate the site-specific row and column index sets. The Cartesian product of these sets indicates the positions in the resultant matrix, $X_{\mathcal{A}^k}^T X_{\mathcal{A}^k}$, where the entries of $X_{\mathcal{A}^k}^{(m)T} X_{\mathcal{A}^k}^{(m)}$ should be placed. We emphasize the structure of the matrix X as it enables the restoration of the variables of interest in the center using local summary statistics. Under this

diagonal structure, $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}$ is sparse in the sense that its remaining entries not filled by the summary statistics are all zeros. The presence of these zero entries eliminates the need for site inter-communication when assembling the resultant matrix. This would be infeasible if a general matrix were used instead.

Algorithm 1 Assemble $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}$ from $X_{\mathcal{A}^k}^{(m)\top} X_{\mathcal{A}^k}^{(m)}$

Require: $|\mathcal{A}^k|, M, X_{\mathcal{A}^k}^{(m)\top} X_{\mathcal{A}^k}^{(m)}$ for $m \in [M]$
Ensure: $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}$
for $m \leftarrow 1 \dots M$ **do**
 Row index set: $R \leftarrow M[|\mathcal{A}^k| - 1] + m$
 Column index set: $C \leftarrow M[|\mathcal{A}^k| - 1] + m$
 $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}_{[R \times C]} \leftarrow X_{\mathcal{A}^k}^{(m)\top} X_{\mathcal{A}^k}^{(m)}$
 $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}_{[(\mathcal{A}^k \setminus R) \times (\mathcal{A}^k \setminus C)]} \leftarrow 0$

Remark If we seek to build $X_{\mathcal{A}^k}^\top X_{\mathcal{I}^k}$, then we replace the column index set C with $M[|\mathcal{I}^k| - 1] + m$.

It is noteworthy that, although the pairwise products required in the center, such as $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}, D^\top D$ is a direct build-up of their local counterparts, their inverse are not. $(X_{\mathcal{A}^k}^\top H X_{\mathcal{A}^k})^{-1}$ and $(D^\top D)^{-1}$ is unattainable from the direct assembling of local summary statistics as every entry of the matrix's inverse requires information across different sites. We, hence, delegate the inverse calculation to the center. Similarly, we can decompose $\alpha^k, d_{\mathcal{I}^k}^k$, the loss L into the pairwise products, which can be directly constructed by local summary statistics sent. However, if ever involved, the inverse will be calculated in the center. Most importantly, all the transmissions we mentioned above only occur once. For example, we only need to transmit $X^\top X$ to the center once. Then, for any \mathcal{A}^k and \mathcal{I}^k , the product of sub-matrices $X_{\mathcal{A}^k}^\top X_{\mathcal{I}^k}$ can be directly obtained from $X^\top X$ stored at the center by using linear algebra such that $X_{\mathcal{A}^k}^\top X_{\mathcal{I}^k} = (X^\top X)_{[\mathcal{A}^k \times \mathcal{I}^k]}$

After overcoming the data privacy constraint, we are ready to introduce Algorithm 2. Following a similar approach to [21], we select potentially important predictors by examining their correlation with the outcome variable y . We initialize set \mathcal{A}^0 by sorting out the corresponding index of T largest $\{\|X_{G_j}^\top y\|_2^2, j \in [p]\}$ in the center, where

$$\|X_{G_j}^\top y\|_2^2 = \sum_{j=1}^M \|X_{G_j}^{(m)\top} y\|_2^2.$$

That requires the local nodes $1, \dots, M$ send $\{\|X_{G_j}^{(m)\top} y\|_2^2, j \in [p]\}$ to the center. It is a one-off transmission.

Algorithm 2 MSplicing (Multi-Site Algorithm)

Require: Local Summary Statistics, C_{max} , \mathcal{A}^0
Ensure: $(\mathcal{A}^{k+1}, \beta^{(m)}, \alpha^{(m)})$, $m = 1, \dots, M$

Initialize $k = 0$.
Compute primal variables β^k , α^k *and dual variable* d^k .

- Call Algorithm 1:
 Assemble the components needed using local summary statistics.
- Solve $\beta_{\mathcal{A}^k}^k \leftarrow (X_{\mathcal{A}^k}^\top H X_{\mathcal{A}^k})^{-1} X_{\mathcal{A}^k}^\top H y$, $\beta_{\mathcal{I}^k}^k \leftarrow 0$;
 $\alpha^k \leftarrow (D^\top D)^{-1} D^\top (y - X \beta^k)$;
 $d_{\mathcal{I}^k}^k \leftarrow X_{\mathcal{I}^k}^\top H (y - X \beta^k) / n$, $d_{\mathcal{A}^k}^k \leftarrow 0$;

while $\mathcal{A}^{k+1} \neq \mathcal{A}^k$ **do**

Record the benchmark loss on the kth iteration.

- Call Algorithm 1:
 Assemble the components needed using local summary statistics.
- Calculate $L^k \leftarrow \frac{1}{2n} \|y - X \beta^k - D \alpha^k\|_2^2$;

Splicing begins:
Select the potential candidates ready for the exchange.

Choose $S_1^k \leftarrow \{j \in \mathcal{A}^k : \sum_{i \in \mathcal{A}^k} I(\|\beta_{G_j}^k\|_2^2 \geq \|\beta_{G_i}^k\|_2^2) \leq C_{max}\}$,
 $S_2^k \leftarrow \{j \in \mathcal{I}^k : \sum_{i \in \mathcal{I}^k} I(\|d_{G_j}^k\|_2^2 \leq \|d_{G_i}^k\|_2^2) \leq C_{max}\}$;

Exchange and try with different exchange-in/out sets sizes.
for $C \leftarrow 1 \dots C_{max}$ **do**
 Let $\tilde{\mathcal{A}}^k \leftarrow (\mathcal{A}^k \setminus S_1^k) \cup S_2^k$, $\tilde{\mathcal{I}}^k \leftarrow (\mathcal{I}^k \setminus S_2^k) \cup S_1^k$;

Recalculate the optimal primary and dual variables values under the new trying set.

- Call Algorithm 1:
 Assemble the components needed using local summary statistics.
- Solve $\tilde{\beta}_{\tilde{\mathcal{A}}^k}^k \leftarrow (X_{\tilde{\mathcal{A}}^k}^\top H X_{\tilde{\mathcal{A}}^k})^{-1} X_{\tilde{\mathcal{A}}^k}^\top H y$, $\tilde{\beta}_{\tilde{\mathcal{I}}^k}^k \leftarrow 0$;
 $\tilde{\alpha}^k \leftarrow (D^\top D)^{-1} D^\top (y - X \tilde{\beta}^k)$;
 $\tilde{d}^k \leftarrow X^\top H (y - X \tilde{\beta}^k) / n$;

Obtain a new loss to be compared against the benchmark.

- Call Algorithm 1:
 Assemble the components needed using local summary statistics.
- Calculate $\tilde{L}^k \leftarrow \frac{1}{2n} \|y - X \tilde{\beta}^k - D \tilde{\alpha}^k\|_2^2$;

If the new loss is significantly smaller, update optimal solutions.
Otherwise, optimal solutions remain unchanged,
but will try other exchanged-in/out set sizes.

if $L^k > \tilde{L}^k$ **then**
 $(\mathcal{A}^{k+1}, \mathcal{I}^{k+1}, \beta^{k+1}, \alpha^{k+1}, d^{k+1}) \leftarrow (\tilde{\mathcal{A}}^k, \tilde{\mathcal{I}}^k, \tilde{\beta}^k, \tilde{\alpha}^k, \tilde{d}^k)$;
break;

else

- Call Algorithm 1:
 Assemble the components needed using local summary statistics.
- Update $S_1^k \leftarrow S_1^k \setminus \arg \max_{i \in S_1^k} \{\|\beta_{G_i}^k\|_{D_{G_i}}^2\}$, $S_2^k \leftarrow S_2^k \setminus \arg \min_{i \in S_2^k} \{\|d_{G_i}^k\|_{D_{G_i}}^2\}$;

Return $(\mathcal{A}^{k+1}, \mathcal{I}^{k+1}, \beta^{k+1}, \alpha^{k+1}, d^{k+1})$;
 Break β^{k+1} , α^{k+1} as $\beta^{(m)}$, $\alpha^{(m)}$, $m = 1, \dots, M$
 send $(\mathcal{A}^{k+1}, \beta^{(m)}, \alpha^{(m)})$ back to local nodes

Remark Note that the input of Algorithm 2 is summary statistics rather than the raw data for data privacy protection concerns.

Allied with the idea of [21], our algorithm solves a constrained optimization problem. There are two major differences between Algorithm 1 and that of [21]. The first difference is that we have rewritten the solutions, loss, and set selection criteria with α unpenalized. The second difference is that, due to the distributed learning setting, we need to expand and break down the expression of pertinent variables pairwise, as demonstrated in Eq. (13), before calculating the variables of interest. We achieved that by calling the Algorithm 1. However, the convenience is evident as these pairwise products are already obtainable in the center, having no demands for additional communication between sites. Being a one-time occurrence, the data transmission safeguards data privacy and enables the adoption of more advanced security techniques, such as differential privacy, in future work.

One may wonder why we employ Euclidian distance for the candidate pre-screening in the while loop but use a newly defined norm during splicing. That is because S_1^k is a subset of \mathcal{A}^k . The metrics we utilized to select S_1^k , hence, are associated with the optimal value of \mathcal{A}^k , which minimizes the loss. However, the new selection criteria could be different if we attempt to move in the direction that lessens the loss change, as shown in Lemma 2.1.

Lemma 2.1 For any $j \in \mathcal{A}^k$, the loss increase to $\mathcal{L}(\beta^k)$ due to the disposal of the j th group is

$$\mathcal{L}(\beta^{\mathcal{A}^k \setminus j}) - \mathcal{L}(\beta^k) = \beta_{G_j}^\top \left(X_{G_j}^\top \left(\frac{I + D(D^\top D)^{-1} D^\top}{2n} \right) X_{G_j} \right) \beta_{G_j},$$

where $\beta^{\mathcal{A}^k \setminus j}$ represents the estimator, which we assign the entries of j th group to be zero.

We, hence, define new selection criteria that appeared in Algorithm 2.

$$\|\beta_{G_j}\|_{D_{G_j}}^2 = \beta_{G_j}^\top \left(X_{G_j}^\top \left(\frac{I + D(D^\top D)^{-1} D^\top}{2n} \right) X_{G_j} \right) \beta_{G_j}, \tag{14}$$

where $(D^\top D)^{-1}$ has already been computed in the center and requires no updates. β_{G_j} and $X_{G_j}^\top D$ are directly attainable from β^k , $X_{\mathcal{A}^k}^\top D$ and $X_{\mathcal{I}^k}^\top D$, respectively. For $j \in \mathcal{A}^k$, $X_{G_j}^\top X_{G_j}$ are extractable from $X_{\mathcal{A}^k}^\top X_{\mathcal{A}^k}$. The logic of defining the $\|d_{G_j}\|_{D_{G_j}}^2$ and ways of computing it with summary statistics is similar.

We end this section by accentuating that Algorithm 2 can automatically address *high dimensionality* and *site heterogeneity*. The simulations in Sect. 4 also demonstrate that it can improve *estimation accuracy*. Furthermore, Algorithm 2 excels in aspects including *computation cost*, *privacy protection*, and *equivalence*, and we give remarks to the last three merits.

- **Computation cost:** We extend the algorithm to multiple sites but only spend a similar computation cost compared to the original single-site algorithm, except for the time spent on assembling the summary statistics in Algorithm 1. That is because the proposed method can be viewed as splicing implemented on a single site, i.e., the center, once the summary statistics are assembled, and thus, the proposed method inherits the merits of computation cost in [21] and possesses polynomial complexity.
- **Privacy protection:** During the data transmission procedure, no individual data were exposed or utilized in the algorithm. The summary statistics we constructed replace their roles and suffice the need for the algorithms to function properly under the distributive setting. No communication has ever occurred between local nodes, thus eliminating the potential collusion. The data are only transmitted once, and its receiving end is exclusive to the central node. The favorable one-time off data transfer in our algorithm, hence, easily allows for incorporating other privacy-preserving techniques, such as differential privacy.
- **Equivalence:** The specially designed block diagonal structure and Algorithm 1 ensure the exact recovery of the associated variables to compute the optimal solutions and enable the solution output from Algorithm 2 to be equivalent. That is, the solution output from Algorithm 2 under the DataSHIELD constraint in a multi-site setting is exactly the same as the one outputted from the whole dataset when the datasets from different sites are pooled together at a single site.

3 Theoretical Properties

In this section, we will provide the assumptions and, hence, develop theorems to justify the validity of our proposed method.

Assumption 1 The random errors ε_i for $i \in [n]$ follow an i.i.d. sub-Gaussian distribution with mean zero. Mathematically, there exists $\sigma > 0$ such that $P(|\varepsilon_i| > x) \leq 2 \exp(-x^2/\sigma^2)$, for all $x \geq 0$.

Assumption 2 Let $0 < \hat{c}_*(T) \leq \hat{c}^*(T) < \infty$,

$$\hat{c}_*(T)\|u\|_2^2 \leq \frac{\|\hat{X}_{\mathcal{A}}u\|_2^2}{n} \leq \hat{c}^*(T)\|u\|_2^2, \forall u \neq 0, u \in \mathbb{R}^{M-|\mathcal{A}|} \text{ with } |\mathcal{A}| \leq T.$$

Assumption 3

$$\hat{\omega}_T\|u\|_2 \geq \frac{\|\hat{X}_{\mathcal{A}}^T \hat{X}_{\mathcal{B}}u\|_2}{n}, \forall u \neq 0, u \in \mathbb{R}^{M-|\mathcal{B}|} \text{ with } |\mathcal{A}| \leq T, |\mathcal{B}| \leq T, \text{ and } \mathcal{A} \cap \mathcal{B} = \emptyset.$$

Assumption 4

$$\hat{\mu}_T = \frac{8\hat{c}^*(T)\left(\frac{(1+\hat{\eta})\hat{\omega}_T}{\hat{c}_*(T)}\right)^2}{(1-\hat{\eta})\left(\hat{c}_*(T) - \frac{\hat{\omega}_T^2}{\hat{c}_*(T)}\right)}$$

$0 \leq \hat{\mu}_T < 1$ and the constant $\hat{\mu}_T$ depends on T for some constant $0 < \hat{\eta} < 1$.

Assumption 5 The minimum group signal is denoted as $\vartheta = \min_{j \in \mathcal{A}^*} \|\beta_{G_j}\|_2^2$ and it satisfies $\frac{TM \log(M(p+1)) \log(\log n)}{n\vartheta} = o(1)$, where T is the support size and M, p, n is the total number of sites, nuisance parameters considered, and observations across sites, respectively.

Assumption 6 $\frac{M|\mathcal{A}^*| \log(p+1) \log(\log n)}{n} = o(1)$ and $\frac{MT_{\max} \log(M(p+1))}{n} = o(1)$, where T_{\max} is the maximum support size.

Assumption 7 The total number of sites $M = o((p + 1)^{\log(\log n)})$

The assumptions listed above are either well presented and argued in the work of [21] or are modest extensions of the former; we consider the transformed matrix \hat{X} , given in Eq. (10), instead of the original design matrix X , to embed the effects of site-specific parameters. Additionally, we fix the group size to be M , a scenario included in the more general setting of [21].

Assumption 1 specifies the distribution for the error but not for the nuisance parameter X . A fixed design for the nuisance parameter avoids the distribution heterogeneity issue in X , which occurs in federated learning. Assumptions 2 and 3 control the angle between the column spaces of X 's subpartitions. Specifically, under Assumption 3, when $\hat{\omega}_T = 0$, the column space of \hat{X}_A is orthogonal to that of \hat{X}_B , implying independence between these partitioned matrices. Assumption 4 regulates the correlation between the groups. We provide a stricter and thus better upper bound, δ_T , for ω_T compared to the one given by [21]. Specifically, we set $\delta_T = (c(2T) - c(2T))/2$. One can then verify that $\omega_T \leq \delta_T$ by the theorem of [42]. Zhang et al. [21] provides one sufficient condition to govern Assumption 4, which is $\delta_T \leq 0.188$. Alternatively, that is $c(2T) \geq 0.812$ and $c(2T) \leq 1.188$. However, note that this condition is sufficient but not necessary. Therefore, it may be possible for one to relax this correlation condition further.

Under these assumptions, Theorem 3.1 proves that, with high probability, Algorithm 2 will not incorrectly filter out truly relevant variables. Theorem 3.2 verifies that cooperated with the GIC criterion, the algorithm can identify the subset of truly relevant groups even when the model size is unknown. Its estimator is the same as the oracle least-squares estimator [21]. Theorem 3.2 implies that, with high probability, the GIC-tuned Algorithm 2 has polynomial complexity. The proof of the theoretical results can be found in the appendix.

Theorem 3.1 (Support Recovery) *Let $(\hat{\mathcal{A}}, \hat{\mathcal{L}}, \hat{\beta}, \hat{\alpha}, \hat{d})$ represent the solutions of Algorithm 2. Under Assumptions (1) to (5). if $T \geq s^*$, we obtain*

$$P(\hat{\mathcal{A}} \supseteq \mathcal{A}^*) \geq 1 - \delta_1 - \delta_2.$$

Particularly, when $T = s^*$,

$$P(\hat{\mathcal{A}} = \mathcal{A}^*) \geq 1 - \delta_1 - \delta_2.$$

Theorem 3.2 (Selection Consistency) *Let $(\hat{\mathcal{A}}, \hat{\mathcal{T}}, \hat{\beta}, \hat{\alpha}, \hat{d})$ represent the solutions of Algorithm 2. Suppose Assumptions (2) to (5) hold with T_{\max} , and Assumptions (1), (6), and (7) also hold. When n is sufficiently large, we obtain*

$$P\left(\min_{\hat{\mathcal{A}} \neq \mathcal{A}^*, \hat{\mathcal{A}} \subseteq \mathcal{S}} \text{GIC}(\hat{\mathcal{A}}) > \text{GIC}(\mathcal{A}^*)\right) \geq 1 - O(p^{-\gamma}),$$

for some constant $0 < \gamma < 1$, where

$$\text{GIC}(\hat{\mathcal{A}}) = n \log L(\hat{\beta}) + M(T + 1) \log(p + 1) \log(\log n),$$

and $L(\hat{\beta})$ is the loss output for a fixed support size T from Algorithm 2.

Theorem 3.3 (Convergence) *In the k th iteration, Algorithm 2 outputs solutions denoted as $(\mathcal{A}^k, \mathcal{T}^k, \beta^k, \alpha^k, d^k)$. Under Assumptions (1) to (4), when $T \geq s^*$, we obtain:*

(i)

$$\left|2nL(\beta^k) - 2nL(\beta^*)\right| \leq \hat{\mu}_T^k \|y\|_2^2.$$

(ii)

$$\mathcal{A}^k \supseteq \mathcal{A}^* \text{ if } k > \log_{\frac{1}{\hat{\mu}_T}} \left(\frac{\|y\|_2^2}{\left(1 - \frac{\hat{\eta}}{2}\right)n \left(\hat{c}_*(T) - \frac{\hat{\omega}_T^2}{\hat{c}_*(T)}\right) \vartheta} \right),$$

with probability at least $1 - \delta_1 - \delta_2$.

4 Simulation

Next, we conduct simulations to verify the merits of Algorithm 2. The results demonstrate that Algorithm 2 achieves nearly zero false positive rates (FPR), improved estimation accuracy, and polynomial computational time, as evidenced by our experiments.

We construct each site matrix $X^{(m)} \in \mathcal{R}^{n_m \times p}$, where $m = 1$ to M by drawing its rows independently from a multivariate Gaussian distribution $\mathcal{MVN}(0_{n_m \times 1}, \Sigma_{p \times p})$. We let p columns of $X^{(m)}$ exponentially correlate. That is, we set the covariance matrix, $\Sigma_{ij} = \rho^{|i-j|}$, and it will reflect the interactions between different groups. To depict the relationship between $X^{(m)}$ and $D^{(m)}$ on the same site, we begin by generating $\dot{D}^{(m)}$ according to the rule that if $\sum_{j \in \mathcal{A}^*} X_{ij}^{(m)} > 1$, then $\dot{D}_i^{(m)} = 1$. Otherwise,

$\dot{D}_i^{(m)} = 0$. When we set $D^{(m)} = \dot{D}^{(m)}$, there is a strong correlation between $D^{(m)}$ and $X^{(m)}$. Allowing more flexibility in adjusting the correlation between $D^{(m)}$ and $X^{(m)}$, we introduce the probability \dot{p} . Formally, we will generate $D^{(m)}$ as follows,

$$D_i^{(m)} = \begin{cases} \dot{D}_i^{(m)} \text{ (strongly correlated with } X_{ij}^{(m)}, j \in \mathcal{A}^*) & \text{with probability } \dot{p} \\ 0 \text{ or } 1 \text{ (random assignment)} & \text{each with probability } \frac{1-\dot{p}}{2} \end{cases}$$

When $\dot{p} = 0$, $D^{(m)}$, and $X^{(m)}$ are independent and have zero correlation.

We then generate the underlying regression coefficients β^* . Since β^* can be obtained by re-allocating entries from $\beta_{G_j}^*$, we will define $\beta_{G_j}^*$ instead. Different from bringing in ρ , which portrays the interaction *across* groups, we construct $\beta_{G_j}^*$, where $j \in \mathcal{A}^*$ and α^* with the intention to reflect the correlation *inside* the group. That is, we assume the same predictor has a correlated coefficient across different sites. For each important group’s coefficient entries, $i = 1$ to M , we let

$$(\beta_{G_j}^*)_i = b_i^j - \frac{1}{M} \sum_{i=1}^M b_i^j,$$

where b_1^j, \dots, b_M^j is obtained by independent draws from $\mathcal{N}(0.5, 0.2)$. Meanwhile, we set $\beta_{G_j}^* = 0_{M \times 1}$ for $j \in \mathcal{T}^*$ trivially from the definition of \mathcal{T}^* . Eventually, we obtain y from

$$y = \text{diag}(X^{(1)}, \dots, X^{(M)})\beta^* + \text{diag}(D^{(1)}, \dots, D^{(M)})\alpha^* + \varepsilon,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1$ to n .

Nuisance parameters are fixed to be $p = 100, n_m = 50, M = 4, \rho = 0.9$, and $\sigma = 4$. We will adapt the same metrics used in [21] except for one w.r.t. to the site-specific parameter α to evaluate the selection of group, and parameter estimation. They are:

- True Positive Rate (TPR): $TP/(TP + FN)$, where $TP := |\hat{\mathcal{A}} \cap \mathcal{A}^*|$ and $FN := |\hat{\mathcal{T}} \cap \mathcal{A}^*|$
- False Positive Rate (FPR): $FP/(FP + TN)$, where $TN := |\hat{\mathcal{T}} \cap \mathcal{T}^*|$ and $FP := |\hat{\mathcal{A}} \cap \mathcal{T}^*|$.

Table 1 Summary of simulation settings

Test	Fixed parameters	Varied parameters	Figures
Density of β	$\dot{p} = 0, \alpha^* = (0.9, 1, 1.1, 1.2)$	$s = 5, 10, 15$	1
Correlation	$s = 10, \alpha^* = (0.9, 1, 1.1, 1.2)$	$\dot{p} = 0, 0.5, 1$	2
Parameter properties	$s = 10, \dot{p} = 0.5$	$\alpha^* = (1, 1, 1, 1),$ $(-1.1, 1.2, -1.3, 1.4),$ $(0, 0, 0, 0)$	3
Support of α^*	$s = 10, \dot{p} = 0.5$	$\alpha^* = (1.1, -1.2, 1.3, -1.4),$ $(1.1, -1.2, 0, 0)$	4

Table 2 Comparison of computational times for MSplicing and abess

Scene	$s = 5$	$s = 10$	$s = 15$
MSplicing	0.87 (0.15)	0.86 (0.14)	0.87 (0.14)
Abess	0.47 (0.08)	0.47 (0.08)	0.48 (0.08)
	$\hat{\rho} = 0$	$\hat{\rho} = 0.5$	$\hat{\rho} = 1$
MSplicing	0.87 (0.14)	0.86 (0.13)	0.88 (0.14)
Abess	0.48 (0.08)	0.47 (0.08)	0.48 (0.08)
	Homogeneity	Heterogeneity	Nullity
MSplicing	0.89 (0.14)	0.87 (0.15)	0.87 (0.14)
Abess	0.49 (0.09)	0.47 (0.08)	0.48 (0.08)
	Same	Similar	
MSplicing	0.87 (0.15)	0.86 (0.14)	
Abess	0.48 (0.09)	0.48 (0.08)	

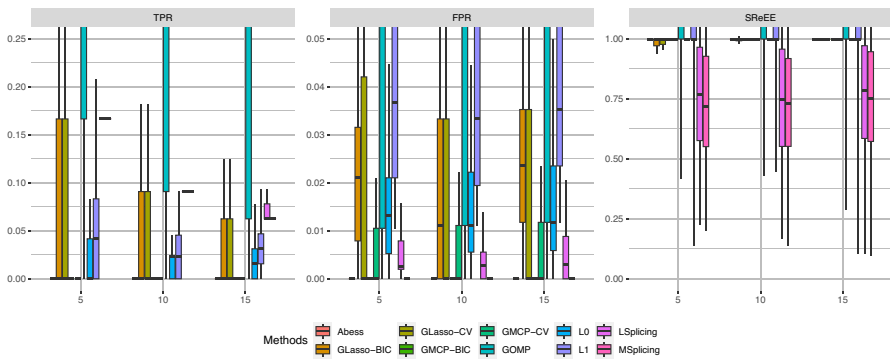


Fig. 1 Test various densities of the nuisance parameters with methods Abess, GLasso-BIC, GLasso-CV, GMCP-BIC, GMCP-CV, GOMP, L0, L1, LSplicing, and MSplicing plotted from left to right in each setting

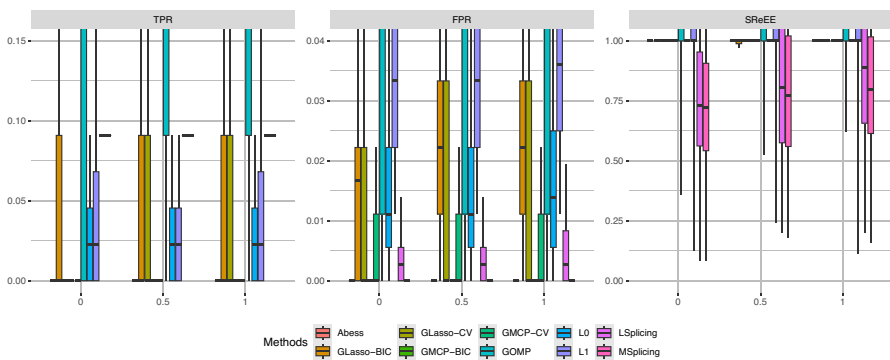


Fig. 2 Test various correlations between X and D with methods Abess, GLasso-BIC, GLasso-CV, GMCP-BIC, GMCP-CV, GOMP, L0, L1, LSplicing, and MSplicing plotted from left to right in each setting

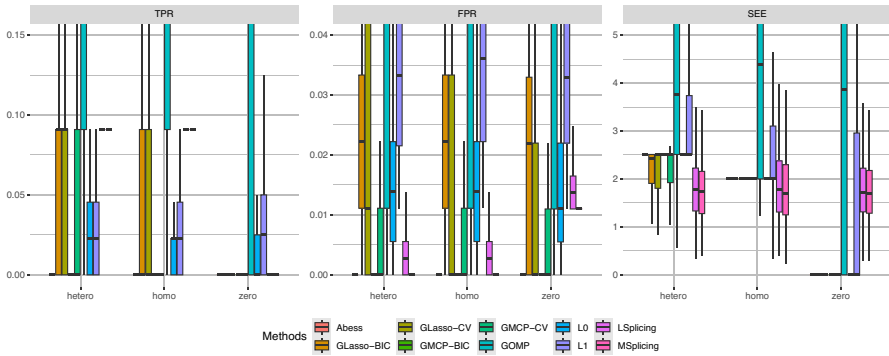


Fig. 3 Test heterogeneity, homogeneity, and nullity of the site-specific parameter with methods Abess, GLasso-BIC, GLasso-CV, GMCP-BIC, GMCP-CV, GOMP, L0, L1, LSplicing, and MSplicing plotted from left to right in each setting

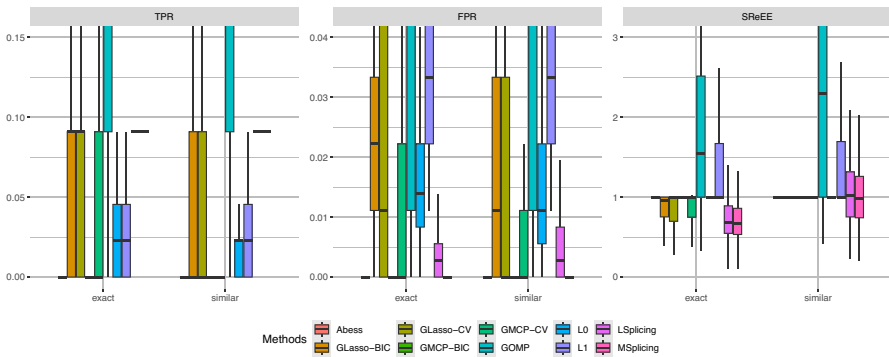


Fig. 4 Test the exact and similar support of the site-specific parameter with methods Abess, GLasso-BIC, GLasso-CV, GMCP-BIC, GMCP-CV, GOMP, L0, L1, LSplicing, and MSplicing plotted from left to right in each setting

- Estimation Error of the site-specific parameter α (SReEE): $SReEE = \|\hat{\alpha} - \alpha^*\|_2 / \|\alpha^*\|_2$ or if the testing setting contains $\|\alpha^*\|_2 = 0$, we use
- Estimation Error of the site-specific parameter α (SEE): $SEE = \|\hat{\alpha} - \alpha^*\|_2$

The closer the TPR to 1 and FPR to 0, the better the method has performed in the feature selection. The lower the SReEE or SEE to 0, the more accurate the site-specific parameter estimation is. Based on the discussion in [21], the methods SGSplicing (Abess), GLasso-BIC (group Lasso using the Bayesian information criterion), GLasso-CV (group Lasso using 5-fold cross-validation), GMCP-BIC (group MCP using the Bayesian information criterion), GMCP-CV (group MCP using 5-fold cross-validation), and GOMP (group orthogonal matching pursuit) will be used as baseline methods. Additionally, we include methods where L_1 or L_0 penalization on both the site-specific parameter and the nuisance parameter is applied at each site, and the parameters are estimated separately based on the local data. Finally, both

the local version of Algorithm 2, i.e., the method applied to the local data, LSplicing, and the proposed method, i.e., the original multi-site version of Algorithm 2, MSplicing, are included.

The sequence of these methods being plotted on the figures follows the order of being presented here. We set the repetition to be 500 times. If a “proper” box plot fails to be seen in the figures, it is because it shrinks into a bold line, indicating a small range of value fluctuations. We will have four settings in total. Each will test the methods’ performance under different conditions: varying densities of β , different correlations between $D^{(m)}$ and $X^{(m)}$, diverse properties of α (heterogeneity, homogeneity, and nullity), and various supports for α (exact and similar). Table 1 provides the summary of different simulation settings.

Figures (1, 2, 3 and 4) show in all settings, our method remains superior in FPR and SReEE. A nearly 0 FPR indicates our approach will not select the wrong features at almost all times, a property inherited from Abess. Abess also stays at almost 0 FPR. In contrast, other methods, such as GOMP, tend to over-choose the predictors and climb to a high FPR, which is undesirable in the feature selection as it leads to the model misspecification problem. Another attribute of our method and Abess shared is its conservativity. It means our algorithm and Abess tend to select fewer predictors to guarantee they will not include the wrong predictors, especially when the noise-signal ratio is high. That explains why both two methods are relatively inactive in the TPR. However, our method performs better each time than Abess in TPR due to the correct use of the prior information, except when we test for α^* with zero coefficients (i.e., the given foreknowledge is wrong). In that scenario, our method is on the same TPR level as Abess. Apart from the case where the existing knowledge is incorrect, our method keeps the lowest α estimation error among all the algorithms. In comparison, other approaches either continue to exhibit a high α error rate or exclude the parameter of interest, α , from the model; SReEE (not SEE) = 1 indicates that some other algorithms have screened out the predictor we wish to retain, which could impair the model’s interpretability. We also notice that, given the prior information is correct, among all three methods run on each local site separately, our local version, LSplicing, outperforms both L_0 and L_1 in every criterion (i.e., TPR, FPR, and SReEE) due to its inclusion of deemed important features. Its superior performance justifies the need for conditional feature selection. Meanwhile, our original multi-site method surpasses its local counterpart, suggesting that employing common support assumptions could further enhance performance.

Because Zhang’s Abess significantly outperformed other algorithms in their paper, we will focus on comparing the computational times of Abess and our method. The result in Table 2 suggests the computational cost of the proposed method is slightly higher than but comparable to the one of Abess. Such an observation aligns with our expectations, as our algorithm is developed from Abess and inherits its polynomial computational complexity, while the assembling of the summary statistics will take some additional time - but it is an inevitable trade-off for privacy protection. Note that here, we report time in R and with parallel computing removed because the time difference will be more obvious in case they are not of the same order compared with that in C.

5 Real Data Applications

The motivation for the proposed method lies in multi-center studies. Here, we consider a single-center real data application in order to better demonstrate our method. In particular, collecting the whole dataset allows us to compare the proposed method with the method without DataSHIELD constraints, such as group lasso. This is a commonly adopted illustration strategy in integrative analysis literature [5]. In the appendix, we include an additional real data analysis on a multi-center energy dataset.

We testify to the aforementioned selected approaches in AIDS Clinical Trials Group Study 175 (ACTG175). We will designate CD8 T cell count at 20 ± 5 weeks (cd820) as the dependent variable and, accordingly, eliminate other outcome variables, including CD4 T cell count at baseline (cd40), at 20 ± 5 weeks (cd420), and 96 ± 5 weeks (cd496), missing CD4 T cell count at 96 ± 5 weeks (r), CD8 T cell count at baseline (cd80), and (days); the number of days until the first instance of (i) a decline in CD4 T cell count of at least 50, (ii) an event indicating progression to AIDS, or (iii) death.

We investigated the treatment effect of zidovudine-incorporated therapy. This means that instead of utilizing the treatment arm (arms) variable with values 0=zidovudine, 1=zidovudine and didanosine, 2=zidovudine and zalcitabine, and 3=didanosine, we employ the treatment indicator (treat). This variable assigns 0 to zidovudine only and 1 to other therapies.

When selecting appropriate covariates, we have excluded the patient ID number (pidnum) due to its lack of predictive power. Additionally, we have removed the dummy variable zidovudine used before treatment initiation (zprior) since it has a constant value 1 across all observations. As our analysis does not involve survival analysis, the indicator variable for observing the event in days (cens) has been deleted. We have retained the remaining 14 predictors, except for the variable race (race). That is because we intend to separate the data into two sites based on ethical value, with 0 representing white and 1 denoting non-white. In such a way, we created a multi-site dataset that mimics data collected from hospitals in different geographical areas. The predictors selected by methods in comparison are as follows.

- Our method: wtkg, karnof, treat
- GLasso-BIC: wtkg, homo, gender
- GLasso-CV: wtkg, hemo, homo, drugs, oprior, preanti, gender, strat, symptom
- GMCP-BIC: No variables are selected
- GMCP-CV: wtkg, hemo, drugs, oprior, preanti, gender, symptom
- GOMP: age, wtkg, hemo, homo, drugs, karnof, oprior, preanti, gender, str2, strat, symptom, offtrt, treat
- SGSplicing: wtkg, gender

Notably, except for GOMP, which selected almost all the predictors, our method stands out by specifically identifying “treat” as a significant predictor when we

intend to perform conditional screening. Furthermore, our method identifies the baseline weight in kilograms (wtkg) as a key predictor. (Wtkg) is selected by all the listed methods except for GMCO-BIC, the latter of which did not select any variables at all. Another significant predictor chosen by our method is the Karnofsky score on a scale of 0-100 (karnof). It is a performance scale index, categorizing the patients according to their functional impairment and disability—the lower the score, the smaller the chance for the patients to survive the most severe disease [43]. Our method suggests that the coefficient of treatment for the white group is 18.5 but 40.4 for the non-white group. This discrepancy may indicate a heterogeneous treatment effect for different ethnic groups. Meanwhile, the positive signs and large magnitude may imply that zidovudine-incorporated therapy can universally and significantly improve the CD8 T cell count at 20 ± 5 weeks, regardless of patients' ethnic information. The computational time for our method and Abess is 1.43 s and 0.83 s, respectively.

6 Summary

In this paper, we consider the problem of selecting a common set of active features (support) given data from multiple sites. Among learning tasks across multiple data centers, there is a tendency to use the same features for data analysis for the convenience of merging analysis results or conducting meta-analysis, which motivates our study. To address this issue, we reformulate the common support selection problem as a $L_{2,0}$ penalization problem. To solve the well-known computational challenge in zero-norm penalization, we adopt a splicing-based algorithm with polynomial time complexity. Two improvements are made compared with the existing set selection method: (i) our selection procedure is conditional on the site-specific parameters, which sufficiently takes prior information into account; (ii) our algorithm satisfies the data-sharing constraint, which avoids the privacy leakage when transferring data across different sites. The simulation results also support the superiority of our proposed method in terms of the error rate of variable selection and estimation accuracy of site-specific parameters. We also apply our proposed method to analyze real data, including ACTG 175 and the electricity consumption of multi-site server rooms (see appendix), to show its practicality.

We focus on the common support assumption, and a natural question to ask is whether we can combine the common support assumption with a similar parameter assumption in integrative analysis. To do that, we need to add an appropriate fusion penalization into our objective function and investigate the theoretical properties of the newly defined objective function. In addition, we mainly investigate and split a single-site study to better illustrate the proposed method. It would be definitely of interest to apply the proposed method to establish robust statistical evidence in a real multi-center study in bioscience, which we leave to future work.

Table 3 Coefficients generated by algorithms

	Room Index	11	12	13	14	17	18
MSpling	Temperature α	6294.26	6671.46	7287.76	7014.95	7311.88	7070.06
	Humidity β	5546.28	5201.39	1734.55	3453.53	4551.42	3935.90
GOMP	Humidity β	2084.77	1895.93	1306.63	1542.17	2166.05	1910.18

Appendices

Theoretical Justification

In this section, we will firstly show the equivalence between problem (8) and problem (10), and then combine the theoretical results in [21] to justify the properties of our proposed method.

Equivalence Between Problem (8) and (10)

Proof The Lagrangian form of Eq. (8) can be expressed as the following unconstrained optimization problem:

$$\min_{\alpha, \beta} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 + \lambda \|\beta\|_{2,0},$$

where λ is related with T . Observe that

$$\min_{\alpha, \beta} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 + \lambda \|\beta\|_{2,0} = \min_{\beta} \min_{\alpha} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 + \lambda \|\beta\|_{2,0}. \tag{15}$$

Since $\frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 + \lambda \|\beta\|_{2,0}$ is convex and twice differentiable w.r.t. α for given β , it is obvious that

$$\begin{aligned} \arg \min_{\alpha} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 + \lambda \|\beta\|_{2,0} &= \arg \min_{\alpha} \frac{1}{2n} \|y - X\beta - D\alpha\|_2^2 \\ &= (D^T D)^{-1} D^T (y - X\beta). \end{aligned}$$

Plug-in this minimizer of α in (15), it yields

$$\begin{aligned} (15) &= \min_{\beta} \frac{1}{2n} \|y - X\beta - D(D^T D)^{-1} D^T (y - X\beta)\|_2^2 + \lambda \|\beta\|_{2,0} \\ &= \min_{\beta} \frac{1}{2n} \|(I - D(D^T D)^{-1} D^T)(y - X\beta)\|_2^2 + \lambda \|\beta\|_{2,0} \\ &= \min_{\beta} \frac{1}{2n} \|\hat{y} - \hat{X}\beta\|_2^2 + \lambda \|\beta\|_{2,0} \end{aligned}$$

which corresponds to the Lagrangian form of problem (10). Hence, we have shown the equivalence. □

Justification for Theorems 3.1 to 3.3

Proof Since the problem we study can be reformulated to the problem considered in [21] via a projection matrix H , and the Assumption (1) to (7) in Sect. 3 also correspond to what was required to justify the results in [21], our proposed estimator would spontaneously enjoy the same theoretical properties as that in [21]. In other words, Theorem 3.1 to 3.3 would hold. \square

Additional Numerical Results

Similarly, the merit of our proposed method can be shown by comparing it with existing methods in analyzing a real multi-center data set provided by the data center company, which includes records of temperature, humidity, and daily electricity consumption from eight of their server centers, and the transmission of such data might not be easy due to legal barriers. For each server center m , we have data $X_{(m)}$, which contains observations of temperature, humidity, and the second through fifth powers of the humidity values. We then scaled the columns related to each power of humidity to make them orthogonal to one another. Based on industry experience, temperature is known to be related to electricity consumption. We deemed it important and decided to retain it for further analysis. We set $y_{(m)}$ as the daily electricity consumption at each site m , and added independent noise following $\mathcal{N}(0, 0.5)$ to each site's value.

The result is that our method selects both temperature and the first power of humidity as significant features. However, GOMP only selects the first power of humidity as a significant feature, which prevents us from analyzing our variable of interest: temperature. Meanwhile, Abess and other approaches (GLasso-BIC/CV, GMCP-BIC/CV) remained conservative and did not select any predictors. Table 3 presents the coefficients generated by our method and GOMP. The positive sign in the coefficients outputted by our method also aligns with the physics principle that lower temperature and humidity can accelerate heat exchange in the air, cool the system faster, and thus reduce electricity consumption. The computational time for our method and Abess is 0.39 s and 0.18 s, respectively.

Acknowledgments We would like to thank Xueqin Wang for his kind help and advice on this project and this research is partially supported by the Seed fund of the Big Data for Bio-Intelligence Laboratory (Z0428) and the grant L0438 from The Hong Kong University of Science and Technology.

Funding Open access funding provided by Hong Kong University of Science and Technology.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P (2019) Machine learning and integrative analysis of biomedical big data. *Genes* 10(2):87
2. Niu B, Yuan X-C, Roeper P, Su Q, Peng C-R, Yin J-Y, Ding J, Li H, Lu W-C (2013) Hiv-1 protease cleavage site prediction based on two-stage feature selection method. *Protein Pept Lett* 20(3):290–298
3. Kim G, Kim Y, Lim H, Kim H (2010) An mlp-based feature subset selection for hiv-1 protease cleavage site analysis. *Artif Intell Med* 48(2):83–89. <https://doi.org/10.1016/j.artmed.2009.07.010>
4. Liu H, Shi X, Guo D, Zhao Z, et al (2015) Feature selection combined with neural network structure optimization for hiv-1 protease cleavage site prediction. *BioMed Res Int*
5. Liu M, Xia Y, Cho K, Cai T (2021) Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *J Mach Learn Res* 22(1):5607–5632
6. Cai T, Liu M, Xia Y (2022) Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *J Am Stat Assoc* 117(540):2105–2119
7. Beckmann JS, Lew D (2016) Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Med* 8:1–11
8. Haidich A-B (2010) Meta-analysis in medical research. *Hippokratia* 14(Suppl 1):29
9. Xu H, Platt RW, Luo Z-C, Wei S, Fraser WD (2008) Exploring heterogeneity in meta-analyses: needs, resources and challenges. *Paediatr Perinat Epidemiol* 22:18–28
10. Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, LaFlamme P, Tobin MD, Macleod J, Little J et al (2010) Datashield: resolving a conflict in contemporary bioscience-performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 39(5):1372–1382
11. Tang L, Zhou L, Song PX (2016) Method of divide-and-combine in regularised generalised linear models for big data. arXiv preprint [arXiv:1611.06208](https://arxiv.org/abs/1611.06208)
12. Lee JD, Liu Q, Sun Y, Taylor JE (2017) Communication-efficient sparse regression. *J Mach Learn Res* 18(1):115–144
13. Battey H, Fan J, Liu H, Lu J, Zhu Z (2018) Distributed testing and estimation under sparse high dimensional models. *Ann Stat* 46(3):1352
14. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, Ohno-Machado L (2015) Webdisco: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 22(6):1212–1219
15. Li W, Liu H, Yang P, Xie W (2016) Supporting regularized logistic regression privately and efficiently. *PLoS ONE* 11(6):0156479
16. Predd JB, Kulkarni SR, Poor HV (2009) A collaborative training algorithm for distributed learning. *IEEE Trans Inform Theory* 55(4):1856–1871
17. Mohri M, Sivek G, Suresh AT (2019) Agnostic federated learning. In: *International Conference on Machine Learning*, pp. 4615–4625. PMLR
18. Li Q, He B, Song D (2021) Model-contrastive federated learning
19. Wang J, Liu Q, Liang H, Joshi G, Poor HV (2020) Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv Neural Inform Process Syst* 33:7611–7623
20. Smith V, Chiang CK, Sanjabi M, Talwalkar AS (2017) Federated multi-task learning. *Adv Neural Inform Process Syst* 30:89
21. Zhang Y, Zhu J, Zhu J, Wang X (2023) A splicing approach to best subset of groups selection. *INFORMS J Comput* 35(1):104–119
22. Tang L, Zhou L, Song PX-K (2020) Distributed simultaneous inference in generalized linear models via confidence distribution. *J Multivar Anal* 176:104567
23. Zhang X, Cheng G (2017) Simultaneous inference for high-dimensional linear models. *J Am Stat Assoc* 112(518):757–768
24. Abadie A, Imbens GW (2016) Matching on the estimated propensity score. *Econometrica* 84(2):781–807
25. Lin Z, Ding P, Han F (2021) Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica* 91(6):2187–2217
26. Hahn J (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 8:315–331
27. Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection amongst high-dimensional controls. *Rev Econ Stud* 81(2):608–650

28. Guo X, Wei W, Liu M, Cai T, Wu C, Wang J (2023) Assessing the most vulnerable subgroup to type ii diabetes associated with statin usage: Evidence from electronic health record data. *J Am Stat Assoc* 6:1–12
29. Mun J, Lindstrom MJ (2013) Diagnostics for repeated measurements in linear mixed effects models. *Stat Med* 32(8):1361–1375
30. Ling Q, Tian Z (2011) Decentralized support detection of multiple measurement vectors with joint sparsity. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2996–2999. IEEE
31. Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, New York
32. Harrar SW, Kong X (2016) High-dimensional multivariate repeated measures analysis with unequal covariance matrices. *J Multivar Anal* 145:1–21
33. Zhong P-S, Lan W, Song PX, Tsai C-L (2017) Tests for covariance structures with high-dimensional repeated measurements
34. Ziniel J, Schniter P (2012) Efficient high-dimensional inference in the multiple measurement vector problem. *IEEE Trans Signal Process* 61(2):340–354
35. Jiang Y, He Y, Zhang H (2016) Variable selection with prior information for generalized linear models via the prior lasso method. *J Am Stat Assoc* 111(513):355–376
36. Cui J, Liu Y, Xu Y, Zhao H, Zha H (2013) Tracking generic human motion via fusion of low-and high-dimensional approaches. *IEEE Trans Syst Man Cybernet Syst* 43(4):996–1002
37. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B* 67(1):91–108
38. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, Santoyo-Lopez J, Dunican DJ, Long A, Kelleher D et al (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* 6(8):569–575
39. Su L, Shi Z, Phillips PC (2016) Identifying latent structures in panel data. *Econometrica* 84(6):2215–2264
40. Guo J, Zhu W (2018) Dependence guided unsupervised feature selection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32
41. Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. *Ann Stat* 4:802–837
42. Li C-K, Mathias R (1999) Inequalities on the singular values of an off-diagonal block of a Hermitian matrix. *J Inequal Appl* 1999(2):192382
43. Apostolopoulou E, Raftopoulos V, Terzis K, Elefsiniotis I (2010) Infection probability score, apache ii and Karnofsky scoring systems as predictors of bloodstream infection onset in hematology-oncology patients. *BMC Infectious Dis* 10(1):1–8

Authors and Affiliations

Ho Yi Alexis Ho¹ · Shuoxun Xu¹ · Xinzhou Guo¹

✉ Xinzhou Guo
xinzhoug@ust.hk

Ho Yi Alexis Ho
hyhoai@connect.ust.hk

Shuoxun Xu
subk@connect.ust.hk

¹ Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, HkSAR, HKG, China