ARTICLE TYPE

Robust nonparametric integrative analysis to decipher heterogeneity and commonality across subgroups using sparse boosting

Zihan Li | Ziye Luo | Yifan Sun

1Center of Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

Correspondence
*Yifan Sun, No. 59 Zhongguancun Street, Haidian District, Beijing, 100872
Email: sunyifan@ruc.edu.cn

Summary

In many biomedical problems, data are often heterogeneous, with samples spanning multiple patient subgroups, where different subgroups may have different disease subtypes, stages, or other medical contexts. These subgroups may be related, but they are also expected to have differences with respect to the underlying biology. The heterogeneous data presents a precious opportunity to explore the heterogeneities and commonalities between related subgroups. Unfortunately, effective statistical analysis methods are still lacking. Recently, several novel methods based on integrative analysis have been proposed to tackle this challenging problem. Despite promising results, the existing studies are still limited by ignoring data contamination and making strict assumptions of linear effects of covariates on response. As such, we develop a robust nonparametric integrative analysis approach to identify heterogeneity and commonality, as well as select important covariates and estimate covariate effects. Possible data contamination is accommodated by adopting the Cauchy loss function, and a nonparametric model is built to accommodate nonlinear effects. The proposed approach is based on a sparse boosting technique. The advantages of the proposed approach are demonstrated in extensive simulations. The analysis of The Cancer Genome Atlas data on glioblastoma multiforme and lung adenocarcinoma shows that the proposed approach makes biologically meaningful findings with satisfactory prediction.

KEYWORDS:
Integrative analysis, robustness, nonparametric modeling, sparse boosting

1 INTRODUCTION

Detection and estimation of the molecular markers associated with disease outcomes or phenotypes is one of the core tasks in biomedicine. Regression models have been extensively used to link biomarkers to a phenotypic trait. Traditionally, all samples are expected to follow the same regression model. However, in the biomedical field, the classical assumption does not always hold. The samples often span multiple subgroups, where subgroups represent, for example, disease subtypes, stages, and demographic characteristics. Different subgroups may vary in underlying etiology and pathogenesis and therefore have distinct relationships between observed markers and a phenotype of interest. However, because all samples in subgroups have the same
disease, it is reasonable to expect that they share some commonalities in the trait-marker association. To detect the important marker more accurately and understand diseases more comprehensively, it is crucial to explore the underlying heterogeneity and unveil the commonality across different subgroups concerning the trait-marker association.

Meta-analysis is the most straightforward method, under which each subgroup is initially analyzed separately, and then summary statistics are pooled across subgroups. However, the data in biomedicine often have "large p, small n" characteristics, with the number of covariates $p$ much larger than the sample size $n$. For example, a typical cancer microarray study profiles $p \sim 10^{3-4}$ genes on $n \sim 10^{1-3}$ samples. The high-dimensionality of data makes the estimation in each subgroup challenging and often leads to unsatisfactory results for each subgroup and the overall meta-analysis. An alternative is to conduct an integrative analysis. In contrast to meta-analysis, the integrative analysis does not develop a separate model for each subgroup/dataset; instead, it builds a joint model to analyze multiple subgroups/datasets simultaneously. Recent studies have shown that the integrative analysis often outperforms meta-analysis with more accurate marker identification and coefficient estimation. Most integrative analyses focus on developing novel modeling methods to promote the accuracy of identification, estimation, and prediction. Recently, a few integrative analysis approaches have been paying special attention to the heterogeneity/commonality identification across subgroups/datasets. For example, fused lasso approach and joint lasso approach apply a fusion type penalty to encourage agreement between subgroup-specific coefficients. One of the authors and collaborators develops novel integrative analysis methods to automatically identify the commonality and difference across multiple subgroups/datasets with respect to regression coefficients of the group of covariates by using penalization and boosting techniques, respectively.

The above integrative analysis approach, although quite successful, still has limitations. First, most if not all of the existing studies lack robust properties, which are essential to accommodating outliers/contamination in the response variables. In biomedical studies, response with outliers/contamination is often encountered and can be responsible for extremely large/small response values, errors in data recording, and others. Specifically, consider The Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) data analyzed in the study (for more details, please refer to Section 4.1). There are 211 deaths during follow-up, with survival time ranging from 0 to 127.5 months (median: 13.96 months). The left panel of Figure 1 shows the scatterplot of survival times, as well as the mean and three times standard deviation. It can be observed that six subjects are out of the three-standard-deviation range, suggesting that the data have some outliers with respect to response. The nonrobust approaches cannot effectively accommodate the outliers/contamination in response and may lead to biased estimation, false marker identification, and inference, suggesting that robust analysis is urgently needed. Robust analysis approaches have been extensively developed to low-dimensional data and shown to have clear advantages over nonrobust approaches. For the high-dimensional
data, there are some robust analysis approaches, but the development is still much limited, especially for the multiple subgroups/datasets\textsuperscript{14}. The second limitation of the existing approaches is that they usually assume that the covariates have linear effects on the response. However, recent biomedical studies have shown that nonlinear covariate effects are present in many cases\textsuperscript{15}, which is confirmed by a preliminary examination of the GBM data. In the right panel of Figure\textsuperscript{11} we present the plot of survival time (response variable) against the expressions of gene CLIC5. The purple and green colors correspond to females and males, respectively, and the lines are generated by using kernel regression. It is clear that the two curves have significantly nonlinear trends, and the trends are different. Hence, it is necessary to relax the linear effect assumptions in modeling to accommodate the possible nonlinear effects of covariates on response.

Motivated by the aforementioned limitations in current studies, we develop a robust nonparametric integrative analysis approach to identify heterogeneity and commonality in markers’ associations with responses, as well as select important covariates, and estimate effects, in prognosis studies with high-dimensional molecular measurements. The proposed approach has some connections with but, more importantly, significantly advances from the existing ones. First, it adopts a Cauchy (aka Lorentzian) loss function (CLF) to accommodate outliers/contamination in response. CLF was first proposed in the computer vision field\textsuperscript{16} and has recently been applied to general supervised and unsupervised learning tasks for high-dimensional data\textsuperscript{17,18}. Compared to alternative robust approaches, including the popular quantile approach and Huber’s approach, CLF’s approach is computationally simpler, which is critical to high-dimensional data. In addition, most current robust methods are developed for a single subgroup/dataset and have relatively limited developments for multiple subgroups/datasets. Second, a nonparametric additive model is assumed to accommodate the possible nonlinear effects of markers on response. Compared to the classical linear model, it offers a higher degree of flexibility in the form of covariate effect and can model the nonlinear covariate effects. The proposed approach is based on sparse boosting\textsuperscript{19}, which demonstrates competitive performances in high-dimensional data analysis compared to penalization and other techniques\textsuperscript{20,21}. In all, this study is warranted by providing a practically useful new approach for exploring heterogeneity and commonality across multiple high-dimensional datasets.

2 | METHODS

2.1 | Data and model settings

Suppose we have $M$ independent datasets. Each dataset may represent a different stage, subtype, or subpopulation of the same disease. The sample size for dataset $m$ is $n^m$, $m = 1, \ldots, M$, and the total sample size is $N = \sum_{m=1}^{M} n^m$. In dataset $m$, denote $Y^m = (y_{1}^{m}, \ldots, y_{n^m}^{m})^\top$ as the response vector and $X^m = [(X_{1}^{m})^\top, \ldots, (X_{n^m}^{m})^\top]^\top \in \mathbb{R}^{n^m \times p}$ as measurement matrix for $p$ covariates (e.g., gene expression, SNPs, proteomics, and so on) with $X_{ij}^{m} \in \mathbb{R}$. Assume that the same set of covariates are measured in $M$ datasets. Consider the following nonparametric additive model:

$$y_{ij}^{m} \sim g \left( \sum_{j=1}^{p} f_{ij}^{m}(X_{ij}) \right),$$

(1)

where $g(\cdot)$ is the known link function, and $f_{ij}^{m}(\cdot)$’s are unknown smooth functions. For identification purposes, we assume that all additive components $f_{ij}^{m}(\cdot)$ are centered, i.e., $\sum_{j=1}^{p} f_{ij}^{m}(X_{ij}) = 0$ for $j = 1, \ldots, p$ and $m = 1, \ldots, M$. Here, the data have been normalized so that there is no intercept in the model (1). If $f_{ij}^{m}(\cdot) = f_{ij}^{m}(\cdot)$, then covariate $j$ represents a commonality shared by the datasets $m_1$ and $m_2$. Otherwise, it represents a heterogeneity.

We expand each component $f_{ij}^{m}(\cdot)$

$$f_{ij}^{m}(\cdot) = \sum_{k=1}^{K} \beta_{jk}^{m} \phi_{k}(\cdot) \equiv \phi(\cdot) \beta_{jk}^{m},$$

using basis functions $\phi(\cdot) = (\phi_{1}(\cdot), \ldots, \phi_{K}(\cdot))$, where $\beta_{jk}^{m} = (\beta_{j1}^{m}, \ldots, \beta_{jk}^{m})^\top$ is the unknown coefficient vector. To satisfy the identification restriction, the basis functions are constrained to have mean values of zero, i.e., $\sum_{j=1}^{p} \phi(X_{ij}) = 0$ for $j = 1, \ldots, p$ and $m = 1, \ldots, M$. In the numerical study, we adopt the normalized B-spline basis function, which has been widely used in published studies\textsuperscript{22,23} and note that other basis functions are also be applicable. Then model (1) can be rewritten as

$$y_{ij}^{m} \sim g \left( \sum_{j=1}^{p} \phi(X_{ij}) \beta_{jk}^{m} \right),$$

(2)
Because $f_{jm}^*(\cdot) = f_{jm}^{\ast}(\cdot)$ is equivalent to $\beta_j^{m_1} = \beta_j^{m_2}$, we can distinguish the heterogeneity and commonality across multiple datasets according to $\beta_j^{m_1}$’s. Specifically, for covariate $j$, if $\beta_j^{m_1} = \beta_j^{m_2}$, then it represents a commonality shared by two datasets or a heterogeneity.

### 2.2 Robust identification and estimation by sparse boosting

We adopt the sparse boosting method to select important covariates, estimate the unknown effects, and more importantly, identify the heterogeneity and commonality in covariate effects across multiple datasets. Boosting methods have become very popular for high-dimensional data problems because of their low computational cost and powerful performance. Although standard boosting, e.g., L2-Boosting, are demonstrated to be effective methods in constructing a sparse model, they tend to select many covariates irrelevant to the response. To this end, sparse boosting is developed to tackle this problem, which can lead to sparser models by imposing a penalty of model complexity. The sparse boosting is originally designed to a single dataset, and has been generalized to multiple datasets. Recently, Sun et al. proposed a new sparse boosting method, namely CD-Boost, to address the heterogeneity/commonality problem. The model they considered, however, is a parametric one and lacks robust properties. As shown in Figure 1, the nonlinear covariate effects and outliers in response are not rare in biomedical data. Therefore, a modified boosting algorithm, called robust nonparametric integrative (RNP-Int) algorithm, is developed, which is summarized in Algorithm 1.

The main difference/improvement of the proposed method is the consideration on robustness. To accommodate outliers/contamination in the response, we adopt the Cauchy loss function (CLF), and its definition is shown as follows

$$\rho_c(x) = \log \left( 1 + \left( \frac{x}{c} \right)^2 \right)$$

with influence function

$$\psi_c(x) = \frac{\partial \rho_c(x)}{\partial x} = \frac{2x}{x^2 + c^2},$$

which is used to measure the effect of changing the point of the sample on the value of the parameter estimation. Here, $c$ is a constant. We demonstrate CLF function and its influence function in Figure C1 (Appendix C). Compared to the least-squares loss with $\rho(x) = x^2$, CLF’s influence function has an upper bound. Although the least-absolute deviation loss with $\rho(x) = |x|$ as well as the Huber loss can alleviate the effect of the large error, their influence functions have no cutoff. In contrast, CLF’s influence function has an upper bound and its values tend to reduce to zero with the increase of the error. Therefore, it can alleviate the influence of a single sample, especially a sample with a large noise, on estimating the residuals. As a result, CLF has less dependence on the distribution of the noise and is more robust to the noise. When $g$ is a linear regression model, for example, we use CLF to penalize the residual terms, which is defined as

$$\rho_c(\beta^m) = \sum_{i=1}^m \log \left( 1 + \frac{\left( y_i^m - \sum_{j=1}^p \phi(X_{ij}^m) \beta_j^m \right)^2}{c^2} \right), m = 1, \ldots, M,$$

where $\beta^m = \left( (\beta_1^m)^T, \ldots, (\beta_p^m)^T \right)^T$. In this study, we also consider right censored survival responses under accelerated failure time (AFT) model. Details on the model settings and loss functions are provided in Appendix A.

Beyond robustness, there are three main distinctions between the proposed and standard sparse boosting algorithms for multiple datasets. First, when determining in each iteration which covariate is to be included in the model, multiple datasets are considered jointly. Second, the increment $\gamma$ in coefficient vectors of $M$ datasets are determined simultaneously. Specifically, for covariate $j$, we divide $M$ datasets into $L_j$ disjoint groups ($G_1, \ldots, G_{L_j}$), where each group $G_i$ has an identical coefficient vector, i.e., $\beta_j^{m_1} = \beta_j^{m_2}$, for any $m_1 \in G_i$ and $m_2 \in G_i$, and different groups have distinct coefficient vectors, i.e., $\beta_j^{m_1} \neq \beta_j^{m_2}$, for any $m_1 \in G_i$ and $m_2 \in G_{l_2}$ with $l_1 \neq l_2$. Some datasets in $G_i(l = 1, \ldots, L_j)$ are allowed to have identical increments in coefficient vectors to maintain commonality between these datasets. More precisely, for each unempty subset $G$ of $G_i(l = 1, \ldots, L_j)$, we compute the optimal common increment $\hat{\gamma}_G$ in coefficient vectors of datasets in $G$ by minimizing the sum of CLFs. Then, we select the subset $G'$ and the corresponding increment $\gamma'$ in their coefficients vectors which minimizes the objective function $F(G, \hat{\gamma}_G; j)$ based on enumeration method. The objective function $F(G, \hat{\gamma}_G; j)$ has three components. The first component is the loss function, which is the sum of CLFs of $M$ datasets. The second component explicitly penalizes model complexity and hence can promote sparsity. In the literature, there are multiple choices for model complexity measures, for example, Akaike information criterion (AIC), Bayesian information criterion (BIC), cross-validation, and minimum description length (MDL).
Here we use a BIC type measure as an example, which has been employed in published studies. The second penalty explicitly quantifies difference across multiple datasets and hence can encourage the commonality of coefficient vectors, or equivalently, additive components, across multiple datasets. This penalty was firstly proposed in [13] and the introduction of this penalty is another distinction of the proposed method. It is proportional to the number of covariates with different coefficient vectors in two distinct datasets, and takes value in [0, 1]. It is minimized when all coefficient vectors are the same across $M$ datasets, that is, $\beta_j^1 = \ldots = \beta_j^M$ for $j = 1, \ldots, p$, and is maximized when all coefficient vectors are different. $\lambda$ is the tuning parameter, which controls the strength of penalization. On the one hand, when $\lambda = 0$, the penalization disappears, then the proposed method degenerates into the standard sparse boosting algorithm for multiple datasets. On the other hand, when $\lambda = \infty$, the proposed method compels $M$ datasets to have identical $p$ coefficient vectors, i.e., $\beta_j^1 = \ldots = \beta_j^M$, for $j = 1, \ldots, p$.

**Algorithm 1 RNP-Int**

1. **Initialization.** Denote $\beta = (\beta^1, \ldots, \beta^M)$, and $\beta^{(t)} = (\beta^{1(t)}, \ldots, \beta^{M(t)})$ as the estimate of $\beta$ in the $t$th iteration. Set $\beta^{(1)} = 0$.
2. for $t = 1, \ldots, T$ do
3. for $j = 1, \ldots, p$ do
4. Denote $L_j$ as the number of unique vectors in $(\beta_j^{1(t)}, \ldots, \beta_j^{M(t)})$. Let $(G_1, \ldots, G_{L_j})$ be a mutually exclusive partition of $\{1, \ldots, M\}$, such that $\beta_j^{m(t)} = \alpha_l$ for all $l \in G_l$, where $\alpha_l$ is the common value for the $\beta_j^{m(t)}$'s from group $G_l$.
5. Define $G = \{G|G \subseteq G_l, l = 1, \ldots, L_j\} \setminus \emptyset$ and $\gamma = (\gamma_1, \ldots, \gamma_K)^T$.
6. Compute

$$\hat{\gamma}_G = \arg\min_{\gamma} \sum_{m \in G} \rho_c (\beta^{m(t)} + A_j \gamma) \quad G \in G,$$

(3)

where $\rho_c(\cdot)$ is Cauchy loss function (details below). Here $A_j = (1_{(j-1)K+1}, \ldots, 1_{jK})$ with $1_q$ is the length-$pK$ vector with the $q$th component equal to 1 and all others equal to 0.
7. To determine optimal subset $G^j$ and increment of coefficient vectors $\gamma^j$, compute

$$\hat{G}(\gamma^j) = \arg\min_{(G, \hat{\gamma}) \in G_1 \cup G_2} F(G, \hat{\gamma}; j) \equiv \sum_{m=1}^{M} \left[ \rho_c (\beta_j^{m(t)} + A_j \hat{\gamma} ) + \frac{\log n^m}{n^m} \sum_{s=1}^{p} \sum_{m \in G} \chi (\beta_j^{m(t)} + \chi(m \in G) \delta_s j \hat{\gamma}_G = 0) \right]$$

(4)

where $\chi(\cdot)$ is the indicator function, and $1_m$ is a $pK \times M$ matrix with $m$th column being 1s and other being zeros. Here $pen_c(\beta) = 1 - \sum_{m \leq m_c} \sum_{s=1}^{p} \frac{\rho_c^2 (\beta_j^{m(t)} - \beta_j^{m})^2}{(M)^{p}}$, and $\lambda$ is the tuning parameter.
8. end for
9. Determine the optimal $j$ as $\hat{j} = \arg\min_{1 \leq j \leq p} \hat{G}(G^j, \gamma^j; j)$.
10. Update $\beta^{(t+1)} = \beta_j^{(t)} + \nu \sum_{m \in G^\hat{j}} 1_m A_j \gamma^j$, where $\nu$ is the step size. It has been suggested that the choice of $\nu$ is not critical as long as it is small [10], and we set $\nu = 0.1$ following the published studies.
11. Compute

$$S(t) = \sum_{m=1}^{M} \left[ \rho_c (\beta_j^{m(t+1)} + \frac{\log n^m}{n^m} \sum_{s=1}^{p} \chi (\beta_j^{m(t+1)} = 0) + \lambda \text{pen}_c(\beta_j^{(t+1)}) \right]$$

12. end for
13. Select the optimal stopping iteration by $t^* = \arg\min_{t} S(t)$.
14. $\beta^{(t^*)} = (\beta_1^{(t^*)}, \ldots, \beta^M(t^*))$ is the final estimator of $\beta$. The strong learner for dataset $m$ is $L_i^m = g(\sum_{j=1}^{p} \phi(X_{ij}^m) \hat{\beta}_j^m) (i = 1, \ldots, n^m)$.

**Tuning parameter.** The proposed method involves two tuning parameters $c$ and $\lambda$. For $c$, Li et al. [13] suggest taking some small values (for example, 0.1, 0.5, 1, and 2). In our numeric study, we find that the results are not sensitive to $c$ and set $c = 1$ to reduce the computation burden. In practice, one may need to examine multiple $c$ values and select the optimal value using data-driven methods. $\lambda$ controls the commonality among components in multiple datasets. We search for the optimal $\lambda$ value using five-fold
cross-validation (CV). Specifically, each dataset is partitioned randomly into five disjoint subsets with equal sizes. We apply a one-dimensional grid search for $\lambda$ with the search range as $\{2^{-6}, 2^{-5}, \ldots, 2^{-1}\}$.

**Realization.** To facilitate data analysis within and beyond this study, we have developed a mathematics code implementing the proposed method and made it publicly available at HTTPS://github.com/sunyifan1984/RNP-Int. Further, we provide a demo for two datasets with continuous and survival outcomes, respectively. The proposed method is computationally affordable. For example, for a simulation example with $m = 3$, $n^m = 100$ ($m = 1, 2, 3$), and $p = 200$, overall analysis can be accomplished within 15 min using a laptop with standard configurations. The computation time spent on tuning parameters can be greatly reduced by parallel computing.

## 3 APPLICATION TO SIMULATION DATA

### 3.1 Data generation procedure

This section aims to assess the performance of the proposed method through a simulation study. Suppose there are $M = 3$ independent datasets with, for simplicity, equal sample size $n^1 = n^2 = n^3 = 100$. $p = 200$ covariates are generated from a multivariate normal distribution with zero means and unit variances. We consider an autoregressive correlation structure where covariates $j$ and $k$ have a correlation coefficient $0.5^{|j-k|}$. In each dataset, there are 3–6 covariates with nonzero additive components. Concerning the heterogeneity/commonality across three datasets, we consider three scenarios. Firstly, one covariate has identical nonzero components in three datasets, three covariates have the same nonzero component in two of the three datasets, and one covariate has different nonzero components in different datasets. Secondly, all datasets have the same set of important covariates and also the same components, and finally, none of the datasets share the same components. The nonzero additive components in three scenarios are listed below.

- **Scenario 1 for components:**
  
  $f_1^1(x) = x$, $f_2^1(x) = 3\sin(x)$, $f_3^1(x) = 6/(1 + e^{-2x}) - 3$, $f_4^1(x) = 2x$;  
  
  $f_5^2(x) = x$, $f_6^2(x) = x^2 - 1$, $f_7^2(x) = 3e^{-x^2/4} - \sqrt{6};$  
  
  $f_8^3(x) = x$, $f_9^3(x) = x^2 - 1$, $f_{10}^3(x) = 3\sin(x)$, $f_{11}^3(x) = 3e^{-x^2/4} - \sqrt{6}$, $f_{12}^3(x) = -2x$;

- **Scenario 2 for components:**
  
  $f_1^m(x) = x$, $f_2^m(x) = x^2 - 1$, $f_3^m(x) = 2\sin(x)$, $f_4^m(x) = 6/(1 + e^{-2x}) - 3$, $f_5^m(x) = 3e^{-x^2/4} - \sqrt{6}$, $f_6^m(x) = 2x$ for $m = 1, 2, 3$;

- **Scenario 3 for components:**
  
  $f_1^1(x) = x$, $f_2^1(x) = x^2 - 1$, $f_3^1(x) = 2\sin(x)$, $f_4^1(x) = 6/(1 + e^{-2x}) - 3$, $f_5^1(x) = 3e^{-x^2/4} - \sqrt{6}$, $f_6^1(x) = 2x$;  
  
  $f_7^2(x) = 2(x^2 - 1)$, $f_8^2(x) = 4\sin(x)$, $f_9^2(x) = -3e^{-x^2/4} + \sqrt{6}$, $f_{10}^2(x) = 4x$;  
  
  $f_{11}^3(x) = 2x$, $f_{12}^3(x) = -\sin(x)$, $f_{13}^3(x) = -6/(1 + e^{-2x}) + 3$, $f_{14}^3(x) = -2x$.

The rest of the components are zeros. In the aforementioned scenarios, each dataset has some nonlinear components. We also examine the scenario where all nonzero components degenerate into linear functions. Specifically,

- **Scenario 4 for components:**
  
  $f_1^1(x) = f_1^2(x) = f_1^3(x) = x$, $f_2^1(x) = f_6^1(x) = 2x$, $f_3^1(x) = 1.5x$;  
  
  $f_4^2(x) = f_5^2(x) = f_6^2(x) = x$, $f_7^2(x) = -x$, $f_8^2(x) = 2x$, $f_{10}^2(x) = -1.5x$;  
  
  $f_1^3(x) = f_2^3(x) = f_3^3(x) = x$, $f_4^3(x) = -2x$, $f_5^3(x) = 2x$, $f_6^3(x) = 3x$,

where each dataset has six important covariates, and share one common important covariate and also the identical component.

We consider continuous data under the additive model and right-censored survival data under the AFT model with the following error distributions: (1) Error 1: $N(0,1)$, (2) Error 2: 70%N(0,1)+30%Cauchy(0,1), and (3) Error 3: Cauchy(0,1). With censored survival data, the censoring times are generated independently from uniform distributions. The censoring distributions are adjusted so that the overall censoring rate is about 20%. In both simulation and real data analysis, we adopt a cubic $B$-spline basis with 6 equally spaced inner knots for each $f_j^m$ and without intercept; thus the dimension $K$ of the $B$-spline basis is 6.
3.2 Alternative approaches

Besides the proposed robust nonparametric integrative (RNP-Int) approach, we also consider the following alternatives for comparison. The nonrobust nonparametric integrative (referred to as the "NRNP-Int") approach has the same modeling framework as the proposed one. However, it adopts nonrobust loss, i.e., least-squares loss, in estimation, selection procedure, and stopping criterion. The robust parametric integrative (referred to as the "RP-Int") approach is the same as the proposed, except that all covariate effects are assumed to be linear. The nonrobust parametric integrative (referred to as "NRP-Int") is the nonrobust counterpart of RP-Int. The robust nonparametric meta (referred to as the "RNP-Meta") approach applies sparse boosting under robust nonparametric modeling to each dataset separately, and then results are combined across datasets. The robust nonparametric pool (referred to as the "RNP-Pool") approach combines all datasets and applies robust nonparametric sparse boosting. Details on the alternatives RNP-Meta and RNP-Pool are provided in Appendix B. We acknowledge that other integrative analysis approaches can be adopted to analyze the simulated data. The above five alternatives are the most direct competitors as they follow a similar framework as the proposed approach. Specifically, by comparing with NRNP-Int and NRP-Int as well, the robustness of the proposed approach can be directly established. The advantages of nonparametric modeling of the proposed approach can be demonstrated by comparing it with RP-Int. The significance of adopting integrative analysis to identify heterogeneity/commonality across datasets is investigated by comparing with RNP-Meta and RNP-Pool.

For each training data, we also generate independent test data with \( n^1 = n^2 = n^3 = 50 \) in the same manner. The train datasets are used to train models, and test datasets are used to measure the prediction accuracy of estimated models. For all six methods, the maximum number of boosting iterations is limited to \( T = 300 \).

3.3 Performance comparison

For the proposed approach and five alternatives, we evaluate the heterogeneity/commonality identification performance by the true positive (TP-ind) and false positive (FP-ind). When computing the number of positives, for a covariate \( j = 1, \ldots, 6 \) in Scenarios 1-3, and \( = 1, \ldots, 9 \) in Scenario 4, we count the number of pairs of distinct datasets where this covariate has the same effect, and then sum the first six (nine) covariates. Specifically, the true number of positives for Scenarios 1-4 equals \( 7, \ 18, \ 0, \ \text{and} \ 8 \), respectively. We adapt the true positive (TP-var) and false positive (FP-var) to evaluate the variable selection performance, where the number of positives is the number of true non zero components. Estimation and prediction are also evaluated. Specifically, estimation is measured by the root mean integrated squared errors (RMISE), which is defined as

\[
\sqrt{\sum_{m=1}^{M} \frac{1}{n^m} \sum_{j=1}^{p^m} \left( \hat{f}^m_j(X^m_{ij}) - f^m_j(X^m_{ij}) \right)^2}.
\]

Prediction is measured by the mean absolute error (MAE) for the continuous outcome, which is defined as

\[
\sum_{m=1}^{M} \frac{1}{n^m} \sum_{i=1}^{n^m} \frac{|y^m_i - \hat{y}^m_i|}{n^m},
\]

and C statistics (Cstat) for the survival outcome. The Cstat is the time-integrated area under the time-dependent ROC and measures the overall risk prediction quality for survival data, of which the values are between 0.5 (random prediction) and 1 (perfect prediction). In each scenario, we calculate the average and standard deviation of the six metrics over 100 independent runs.

For data with a continuous outcome, the summary results for Scenarios 1-2 and 3-4 are reported in Table 1 and 2 respectively. The results for right-censored survival data are shown in Table C1-C2 (Appendix C). In general, the proposed RNP-Int performs the best in the identification of heterogeneity/commonality in the majority of the scenarios. Without data contamination, RNP-Int and the nonrobust counterpart NRNP-Int have favorable performances. For example, in Scenario 1, under error 1, RNP-Int and NRNP-Int have (TP-ind, FP-ind)=(6.75, 1.60) and (6.05, 3.60), respectively, compared to (6.60, 4.20) for RP-Int, (6.05, 3.60) for NRP-Int, (1.00, 0.30) for RNP-Meta, and (7.00, 11.00) for RNP-Pool. When data have contamination, RNP-Int performs significantly better than the alternatives. For example, in Scenario 1, under error 2, RNP-Int has (TP-ind, FP-ind)=(6.55, 1.80), which is superior to the alternatives: (5.60, 5.20) for NRNP-Int, (6.55, 4.20) for RP-Int, (5.80, 6.00) for NRP-Int, (1.00, 0.75) for RNP-Meta, and (7.00, 11.00) for RNP-Pool. It is also worth noting that identification results by the robust methods generally have smaller standard deviations. Comparison with the robust parametric integrative method, RP-Int, also shows the merits of the proposed method. When some of the components are nonlinear functions (Scenarios 1-3), RP-Int identifies fewer true positives and (or) more false positives. As expected, RP-Int has superior performance when all of the components are linear functions (Scenario 4). It is noted that RNP-Int still has competitive performance even in the worst case, thus providing a safe choice in practice. RNP-Meta and RNP-Pool have the worst performance among all the methods. Specifically, RNP-Meta can hardly identify commonality. RNP-Pool assumes that all datasets follow the same model. As a result, it fails when there are large differences across datasets (Scenarios 1, 3, and 4).
In identifying important covariates, the proposed RNP-Int again has the best or close to the best performance compared with alternatives. In general, RNP-Int can identify more true positives and fewer false positives. The advantage of RNP-Int against its nonrobust counterpart NRNP-Int becomes more pronounced when data have contamination (errors 2 and 3). When all components are linear functions (Scenario 4), the parametric integrative method, RP-Int, has good performance. However, as expected, its performance deteriorates significantly when some components become nonlinear. Across all simulation scenarios, RNP-Meta performs worse than the proposed RNP-Int. Among the robust methods, RNP-Pool has the worst performance when multiple datasets have completely different components (Scenario 3), which is as expected. Although its performance improves with the increase of the level of commonality across datasets, it is still inferior to RNP-Int when there are some differences between datasets.

In terms of estimation and prediction, the proposed RNP-Int method, in general, has better performance. For example, in Scenario 1, under error 3, RNP-Int has smaller estimation and prediction errors (3.08, 13.55) than NRNP-Int (196.11, 71.16), RP-Int (4.42, 14.34), NRNP-Int (202.38, 32.59), RNP-Meta (10.26, 17.75), and RNP-Pool (11.20, 17.75). To visualize the estimation results better, we show the estimation of components under Scenario 1, error 3 (Figure C2 in Appendix C). It can be observed that the proposed approach provides a more accurate estimation. For data with a survival outcome (Table C1 in Appendix C), the overall observed patterns are similar, where the proposed approach has a favorable performance in identification, estimation, and prediction.

In the aforementioned simulation, all datasets have the same number of samples. We also examine the scenario where different datasets have different sample size. Consider the case with $n^1 = 130$, $n^2 = 110$, and $n^3 = 60$. The other settings are the same as described above. We use Scenario 1 as a representative and present results in Table C3 (Appendix C) for data with a continuous outcome and Table C4 (Appendix C) for data with a right-censored survival outcome. The proposed RNP-Int is again observed to have favorable performance in identification of heterogeneity/commonality, variable selection, estimation, and prediction.

4 | APPLICATION TO TCGA DATASETS

As a cancer omics program initiated by the National Institute of Health (NIH), TCGA publishes high-quality profiling data on multiple cancer types. The rich data provide a valuable opportunity to uncover the underlying genetic factor of cancers, and more importantly, to identify the heterogeneity and commonality in the effects of genes on outcomes across different stages, subpopulations, or subtypes of the same type of cancer. In this section, we analyze TCGA data on glioblastoma multiforme (GBM) and lung adenocarcinoma (LUAD). The genetic and clinical data are downloaded from the cBioPortal website (http://www.cbioportal.org/) via the cgdrsc package. We use the processed level-3 gene expression data. Please refer to literature 29 for detailed information on the generation and processing of gene expression data.

4.1 | GBM data

Glioblastoma multiforme (GBM) is the most common primary brain tumor of adults, as well as the most malignant. The response of interest is overall survival, which is right-censored. In the literature, it has been suggested that male patients have pathogenesis different from that of females 30, which is partially confirmed by our preliminary analysis (see Figure 1). However, there is a lack of attention to the difference in some of the existing studies. As such, we divide the patients into two groups: male and female. After excluding subjects with missing in survival outcomes and gene expression, the male group has 158 samples with 133 deaths, and the female has 99 samples with 78 deaths. These two groups are expected to have both heterogeneity (as suggested by the literature) and commonality (as the two groups suffer from the same type of cancer). A total of 16, 270 RNAseq gene expressions are initially available for the analysis. To improve the reliability of the analysis results, we reduce the gene set to the top 200 genes most correlated with the survival time by applying the nonparametric independence screening 31 with normalized cubic B-spline as the basis function. It is noted that this step, which excludes the most irrelevant genes, is a common practice for analyzing microarray data in cancer-related studies, as the number of cancer-related genes is not expected to be large. We rescale the gene expressions (covariates) such that their values have zero means and unit variances.

The nonlinear effect of some genes on the survival time has been observed in Figure 1, which motivates the nonlinear modeling of gene expressions. We adopt the proposed RNP-Int method for this data under the AFT model, of which the tuning parameter $\lambda$ is selected by five-fold CV, in the same way as in simulation. Results are briefly presented in Figure C3 (Appendix C). With RNP-Int, 13 genes are identified. Among them, three genes behave the same for both males and females. Figure 2 presents the estimated
FIGURE 2 Analysis of GBM data using the proposed RNP-Int approach: estimated effects of 13 genes. Blue solid lines correspond to females, and orange dotted lines to males.

gene effects. It is observed that most of the estimated effects diverge from a straight line, suggesting that the nonparametric modeling is more appropriate to describe the gene effects on survival time in our data. A quick literature search suggests that the identified genes have important biological implications. For example, gene ZIC3 is identified for two subgroups. It is a positive regulator of host gene ARRB1 of miR-326, a GBM downregulated miRNA, which may partly lead to the downregulation of ARRB1 and miR-326 in GBM\textsuperscript{32}. We also identify eight genes that are associated with the response in females but not males. Among them, gene ADAM29 has been found to be a GBM-specific gene in the de novo pathway of primary GBM, and it has been suggested as a viable target for the treatment of GBM\textsuperscript{33}. Another gene that we find to be associated with survival time in females only is gene HMG2, which is a transcriptional modulator that mediates motility and self-renewal in normal and cancer stem cells. It has been found that the level of HMGA2 expression increases in the majority of primary human GBM tumors compared to a normal brain, as well as in CD133+ GBM neurosphere cells compared to CD133- cells\textsuperscript{34}. Further, gene HMGA2 is strongly associated with survival in GBM, suggesting it as a promising biomarker for predicting the survival time of GBM patients\textsuperscript{35}. Genes that have significant effects in males but not females include MYOCD and SMS. Gene MYCOD encodes a nuclear protein, which is expressed in the heart, aorta, and smooth muscle cell-containing tissues. It may play a crucial role in cardiogenesis and differentiation of the smooth muscle cell lineage. Gene SMS, a polyamine biosynthetic enzyme, has been reported to be over-expressed in colorectal cancer\textsuperscript{36}. 
Data are also analyzed using alternative methods. The comparative results are provided in Table 3 and 4. It is noted that different methods generate different results. NRNP-Int, RP-Int, and NRP-Int identify 16, 12, and 15 genes, respectively. RNP-Meta identifies 18 genes, and all genes with nonzero effects are identified as behaving differently across two datasets. RNP-Pool identifies five genes. With its particular property, the effects of the identified genes are the same in the two datasets. Fewer overlaps between the proposed RNP-Int and RP-Int, as well as NRP-Int, in the set of identified genes in two datasets, suggest that genes have non-negligible nonlinear effects on survival times. Omitting these can significantly affect the identification results. Detailed estimation results using these alternative methods are provided in Figures C4-C8 (Appendix C).

Based on real data, the true set of important genes is unknown and, thus, it is challenging to evaluate the identification and estimation accuracy directly. To verify the results, we now evaluate the prediction. Specifically, each dataset is randomly partitioned into a training set and a testing set, at a ratio of 4:1. The penalty parameters for the proposed RNP-Int and alternatives NRNP-Int, RP-Int, and NRP-Int are selected by five-fold CV within the training sample. Estimation (and tuning parameters) is conducted with the training set, and prediction is made with the testing set. With the survival data, we compute the log-rank statistics. Higher log-rank statistics indicate better prediction performance. The average values of log-rank statistics over 100 random replicates are 7.783 (RNP-Int), 4.479 (NRNP-Int), 6.559 (RP-Int), 4.736 (NRP-Int), 4.921 (RNP-Meta), and 5.239 (RNP-Pool), respectively. The proposed approach has the highest log rank and thus has the best prediction performance.

4.2 | LUAD data

Lung adenocarcinoma (LUAD) is a form of non-small cell lung cancer, which is the most common type of lung cancer. In our analysis, we are interested in the regulation of FEV1 (forced expiratory volume in 1 second, prebroncholiator), a critical measure of lung function, by gene expressions. The left panel of Figure C9 (Appendix C) shows the scatterplot of FEV1. It can be observed that there are indeed some outliers that are outside the three-standard-deviation range. The right panel of Figure C9 (Appendix C) presents the plot of FEV1 against the expression of gene ARHGAP4, as well as regression curves and bootstrap-based 95% confidence intervals. It is easy to see that, for different stages of LUAD, the effects of genes on FEV1 are significantly different, and the effects are likely to be nonlinear. However, some of the existing analyses use samples of all tumor stages with insufficient attention to the potential heterogeneity across different stages and also neglect the possible nonlinear effects of gene expression on the response variable. After removing missing values from the FEV1, a total of 225 patients, with 139 in stage 1, 52 in stage II, and 34 in stages III and IV are included in the analysis. A total of 18, 325 RNAseq gene expression measurements are available. Using the same marginal screening process as described above, the number of gene expressions is reduced to 200. All gene expressions are rescaled to a range of 0-1.

Under optimal tuning, the proposed RNP-Int method identifies 17 genes in at least one dataset. Among them, 12 genes behave the same across the three stages, and the rest behave differently. The estimated gene effects are shown in Figure C10 (Appendix C). We again observe that most of the estimated effects may not have a simple linear approximation. A literature review suggests that the findings are biologically meaningful. For example, gene IL31 is identified in all three stages. It has been known that IL31 has important effects on the pathogenesis of allergic diseases and tumors. One genetic variant of IL31, rs4758680, is found to be significantly associated with metastasis and poor survival status in LUAD. ADIPOQ is a protein-coding gene. The encoded protein circulates in the plasma and is involved with metabolic and hormonal processes. The TT genotype of rs2241766, a single-nucleotide polymorphism (SNP) in the gene ADIPOQ, is significantly associated with susceptibility to non-small cell lung cancer and is also related to the overall survival of non-small cell lung cancer patient. In addition to the genes associated with FEV1 in all three stages, we have also identified genes that have strong effects only in the earlier or later stages. Gene ARHGAP4 is found to have strong effects only in stages III and IV, while gene DCLK2 is found to be stage I specific. Genes that are associated with FEV1 in later stages are FRG2C and CHRM3. Gene FRG2C has been found in the non-ground-glass opacity LUAD, and it has been indicated to have significant biological functions related to the cell cycle and proliferation. Gene CHRM3 controls smooth muscle contraction and its stimulation causes secretion of the glandular tissue. The chemotherapy drug docetaxel resistant human NSCLC (A549) cell shows increased expression of CHRM3.

We also apply the alternative methods described above to this data. The summary comparison results in Table 3 and 4 again suggest that different methods produce diverse results. Detailed estimation results using the alternatives are provided in CT15 (Appendix C). Prediction is evaluated as described above. With the continuous data, we compute the mean absolute error (MAE). A lower MAE indicates better prediction performance and the average MAE over 100 random replicates are 2.381 (RNP-Int), 2.867 (NRNP-Int), 2.547 (RP-Int), 2.960 (NRP-Int), 3.114 (RNP-Meta), and 2.406 (RNP-Pool), respectively. The proposed approach is again observed to have better prediction performance.
5 | CONCLUSIONS

In the current biomedical field, collected samples often span multiple subgroups. It is an important yet challenging problem to unveil the underlying heterogeneity and commonality across these related but different subgroups. In this paper, we propose a novel integrative analysis approach called RNP-Int for high-dimensional regression under multiple subgroups/datasets settings. RNP-Int advances from the existing literature by adopting a Cauchy loss function to accommodate possible data contamination and builds a nonparametric additive model to describe the effects of covariates in a more flexible manner. The proposed approach follows the sparse boosting framework, which has an appealing performance in high-dimensional data analysis. Significantly advanced from most of the existing sparse boosting approaches for multiple datasets, the proposed RNP-Int can identify heterogeneity and commonality across multiple datasets, as well as adopt a robust loss function and robust criteria. Simulation has demonstrated that RNP-Int outperforms five direct competitors in the identification of heterogeneity and commonality, variable selection, effect estimation, and prediction. In the analysis of two TCGA datasets, RNP-Int has led to sensible findings, which are significantly different from the alternatives.

In this study, we have adopted the Cauchy loss function to accommodate possible data contamination. Multiple robust methods have been developed for high-dimensional data. However, no method is particularly better than others are. Recently, Barron proposed a general robust loss function, which is a superset of many common robust loss functions, for example, Huber, least-squares, and Cauchy loss functions. By introducing robustness as a continuous parameter, this loss function can adaptively determine its specific form during the training process, which is particularly effective in real data modeling. It would be interesting to extend the current robust loss function with fixed form to this more general and adaptive robust loss function. Beyond the additive model considered in this study, there are many other nonparametric models. It will also be of interest to extend the proposed analysis to other more complex nonparametric models, for example, single- and multi-index models. Finally, it has been recognized that not only genes, but also gene-gene interactions play fundamental roles in understanding, modeling, and treating complex diseases. Interaction analysis has been conducted in genetic studies, but most of the existing attention has been on a single dataset. The integrative interaction approach with special attention to identifying heterogeneity/commonality across multiple datasets is still limited and expected to be very challenging. As considerable new developments will be needed, we postpone this interesting extension to future research.

ACKNOWLEDGMENTS

This study was supported by the Fund for Building World-class Universities (disciplines) of Renmin University of China (18XNB004). The computer resources were provided by the Public Computing Cloud Platform of Renmin University of China.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Curis C, Shah S, Chin S, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; 486(7403): 346-352.
2. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 2015; 21(11): 1350-1356.
3. Guerra R, Goldstein D. *Meta-analysis and Combining Information in Genetics and Genomics*. Boca Raton, FL: Chapman and Hall/CRC. 2009.
4. Yang P, Yang YH, Zhou BB, et al. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics* 2010; 5(4): 296-308.
5. Huang Y, Huang J, Shia B, et al. Identification of cancer genomic markers via integrative sparse boosting. *Biostatistics* 2012; 13(3): 509–522.
6. Huang Y, Zhang Q, Zhang S, et al. Promoting similarity of sparsity structures in integrative analysis with penalization. *Journal of the American Statistical Association* 2016; 112(517): 342–350.

7. Huang Y, Liu J, Yi H, et al. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. *Statistics in Medicine* 2017; 36(3): 509–559.

8. Li Y, Li R, Lin C, et al. Penalized integrative semiparametric interaction analysis for multiple genetic datasets. *Statistics in Medicine* 2019; 38(17): 3221-3242.

9. Tang L, Song PX. Fused lasso approach in regression coefficients clustering–learning parameter heterogeneity in data integration. *Journal of Machine Learning Research* 2016; 17:1-23.

10. Dondelinger F, Mukherjee S. The joint lasso: high-dimensional regression for group structured data. *Biostatistics* 2020; 21(2): 219-235.

11. Sun Y, Jiang Y, Li Y, et al. Identification of cancer omics commonality and difference via community fusion. *Statistics in Medicine* 2019; 38(7): 1200-1212.

12. Sun Y, Sun Z, Jiang Y, et al. An integrative sparse boosting analysis of cancer genomic commonality and difference. *Statistical Methods in Medical Research* 2020; 29(5): 1325-1337.

13. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). *Practical Assessment Research and Evaluation* 2010; 9(6): 1-12.

14. Filzmoser P, Nordhausen K. Robust linear regression for high-dimensional data: An overview. *WIREs Computational Statistics* 2021; 13(4): e1524.

15. Wu M, Ma S. Robust semiparametric gene-environment interaction analysis using sparse boosting. *Statistics in Medicine* 2019; 38(523): 4625-4641.

16. Black MJ, Anandan P. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding* 1996; 63(1): 75-104.

17. Li X, Lu Q, Dong Y, et al. Robust subspace clustering by the Cauchy loss function. *IEEE transactions on neural networks and learning systems* 2019; 30(7): 2067-2078.

18. Barron JT. A general and adaptive robust loss function. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019: 4326-4334

19. Buhlmann P, Yu B. Sparse boosting. *Journal of Machine Learning Research* 2006; 7: 1001–1024.

20. Yue M, Li, J, Ma, S. Sparse boosting for high-dimensional survival data with varying coefficients. *Statistics in Medicine* 2018; 37(5): 789-800.

21. Wu M, Ma S. Robust semiparametric gene-environment interaction analysis using sparse boosting. *Statistics in Medicine* 2019; 38(6): 4625-4641.

22. Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. *The Annals of Statistics* 2010; 38(4): 2282-2313.

23. Li Y, Wang F, Li R, et al. Semiparametric integrative interaction analysis for non-small-cell lung cancer. *Statistical Methods in Medical Research* 2020; 29(10): 2865-2880.

24. Buhlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science* 2007; 22(4): 477–505.

25. Buhlmann P, Yu B. Boosting. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010; 2(1): 69–74.

26. Bühlmann P, Bin Y. Boosting with the l2-loss: regression and classification. *Journal of the American Statistical Association* 2003; 98(462): 324-339.
27. Huang Y, Liu J, Yi HD, et al. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. *Statistics in Medicine* 2017; 36(3): 509–559.

28. He X, Wang SG. Breakdown points of t-type regression estimators. *Biometrika* 2000; 87(3): 675-687.

29. Dewey CN, Li B. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* 2011; 12(1): 323-323.

30. Ostrom Q, Rubin J, Lathia J, et al. Females have the survival advantage in glioblastoma. *Neuro-oncology* 2018; 20 (4): 576-577.

31. Fan J, Feng Y, Song Y. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* 2011; 106(494): 544-557.

32. Nawaz Z, Patil V, Paul Y, et al. PI3 kinase pathway regulated miRNome in glioblastoma: identification of miR-326 as a tumor suppressor miRNA. *Molecular Cancer* 2016; 15(1): 74.

33. Rahane C, Kutzner A, Heese K. A cancer tissue-specific FAM72 expression profile defines a novel glioblastoma multiform (GBM) gene-mutation signature. *Journal of Neuro-Oncology* 2018; 141(1): 57-70.

34. Kaur H, Ali S, Huey L, et al. The transcriptional modulator HMGA2 promotes stemness and tumorigenicity in glioblastoma. *Cancer letters* 2016; 377(1): 55-64.

35. Liu Y, Shete S, Etzel C, et al. Polymorphisms of LIG4, BTBD2, HMGA2, and RTE1 genes involved in the double-strand break repair pathway predict glioblastoma survival. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 2010; 28(14): 2467-74.

36. Guo Y, Ye Q, Deng P, et al. Spermine synthase and MYC cooperate to maintain colorectal cancer cell survival by repressing Bim expression. *Nature Communications* 2020; 11(1): 3243.

37. Yang Y, Li L, Chen F, et al. The role of Interleukin-31 polymorphisms in non-small cell lung cancer genetic susceptibility and clinical outcome. *Genetic testing and molecular biomarkers* 2018; 22(5): 314-319.

38. Cui E, Deng A, Wang X, et al. The role of adiponectin (ADIPOQ) gene polymorphisms in the susceptibility and prognosis of non-small cell lung cancer. *Biochemistry and Cell Biology* 2011; 89(3): 308-13.

39. Kim D, Lee JY, Yoo JY, et al. Genetic features of lung adenocarcinoma with ground-glass opacity: what causes the invasiveness of lung adenocarcinoma?. *The Korean Journal of Thoracic and Cardiovascular Surgery* 2020; 53(5): 250-257.

40. Bland T, Wu B. Repurposing antimuscarinics to treat lethal chemoresistant prostate and lung cancers. *Journal of Clinical Oncology* 2020; 38(15): e15636-e15636.
APPENDIX

A ACCELERATED FAILURE TIME MODEL

The AFT model is an alternative to the commonly used Cox model in survival analysis because of its simple form, lucid interpretation, and low computational cost. In dataset \( m \), for the \( i \)th subject, denote \( T_i^m \) as the survival time and \( C_i^m \) as the censoring time. In the AFT model setting, the nonparametric additive model is specified as
\[
\log(T_i^m) = \sum_{j=1}^p \phi(X_{ij})\beta_j^m + e_i^m,
\]
where \( e_i^m \) is the random error with an unknown distribution function. We assume that data have been properly normalized so that the intercept can be omitted. Under right censoring, we observe \( \{(y_i^m, \delta_i^m, X_{ij}^m), i = 1, \ldots, n^m, m = 1, \ldots, M \} \) where \( y_i^m = \log(\min(T_i^m, C_i^m)) \), \( \delta_i^m = 1(T_i^m \leq C_i^m) \), and \( X_{ij}^m \) are the covariates associated with \( y_i^m \). Assume that \( (y_i^m, \delta_i^m, X_{ij}^m) \)'s have been ordered according to \( y_i^m \)'s in an ascending order.

We adopt the Kaplan-Meier weights to accommodate censoring. The Kaplan-Meier weights \( \omega_i^m \) can be computed as
\[
\omega_i^m = \frac{\delta_i^m}{n^m},
\]
\[
\omega_i^m = \frac{\delta_i^m}{n^m - i + 1} \prod_{k=1}^{i-1} \left( \frac{n^m - k}{n^m - k + 1} \right)^{\delta_k^m}, i = 2, \ldots, n^m.
\]

To accommodate contamination in survival time, we adopt the weighted Cauchy loss function:
\[
\rho_c(\beta^m) = \sum_{i=1}^{n^m} \omega_i^m \log \left[ 1 + \left( y_i^m - \sum_{j=1}^p \phi(X_{ij}^m)\beta_j^m \right)^2/c^2 \right], m = 1, \ldots, M.
\]
ALTERNATIVE APPROACHES RNP-META AND RNP-POOL USED IN NUMERICAL STUDY

RNP-Meta and RNP-Pool are two non-integrative approaches. RNP-Meta conducts sparse boosting under robust nonparametric modeling on each dataset separately, and then results are combined across datasets; RNP-Pool combines all datasets directly and applies the sparse boosting under robust nonparametric modeling. The algorithms of RNP-Meta and RNP-Pool are summarized in Algorithm 2 and 3.

Algorithm 2 RNP-Meta

1: Initialization. Set $\beta^{(1)} = 0$.
2: for $m = 1, \ldots, M$ do
3:     for $t = 1, \ldots, T$ do
4:         compute $(\hat{s}^m, \hat{y}^m) = \arg\min_{1 \leq s \leq p, \gamma \in \mathbb{R}^K} \rho_{\varepsilon} \left( \beta^{(t)} + \sum_{k=1}^{K} \gamma_k 1_{(s-1)K+k} \right) + \frac{\log n^m}{n^m} \sum_{j=1}^{p} \chi \left( \hat{\beta}^{(t)}_j + \delta_j \gamma \neq 0 \right)$.
5:     Update $\beta^{(t+1)} = \beta^{(t)} + \nu \sum_{k=1}^{K} \hat{y}^m_k 1_{(s-1)K+k}$, where $\nu$ is the step size as in Algorithm 1.
6:     Compute $F^{(t)} = \rho_{\varepsilon}(\beta^{(t+1)}) + \frac{\log n^m}{n^m} \sum_{j=1}^{p} \chi \left( \hat{\beta}^{(t+1)}_j \neq 0 \right)$.
7: end for
8: Select the optimal stopping criterion by $\hat{t}^m = \arg\min_{1 \leq t \leq T} F^{(t)}$.
9: $\hat{\beta}^{(\hat{t})}$ is the final estimator of $\beta^m$. The strong learner for dataset $m$ is $L^m = g \left( \sum_{j=1}^{p} \phi(X^m_i) \hat{\beta}^{(\hat{t})}_j \right)$ with $i = 1, \ldots, n^m$.
10: end for

Algorithm 3 RNP-Pool

1: Initialization. Combine $M$ datasets into one, and denote $Y = [(Y^1)^\top, \ldots, (Y^M)^\top]^\top \in \mathbb{R}^N$ as the combined responses, $X = [(X^1)^\top, \ldots, (X^M)^\top]^\top \in \mathbb{R}^{N \times p}$ as the combined measurement matrix, and $\beta \in \mathbb{R}^{Kp}$ as the unknown coefficient vector. Set $\beta^{(1)} = 0$.
2: for $t = 1, \ldots, T$ do
3:     compute $(\hat{s}, \hat{y}) = \arg\min_{1 \leq s \leq p, \gamma \in \mathbb{R}^K} \rho_{\varepsilon} \left( \beta^{(t)} + \sum_{k=1}^{K} \gamma_k 1_{(s-1)K+k} \right) + \frac{\log N}{N} \sum_{j=1}^{p} \chi \left( \hat{\beta}^{(t)}_j + \delta_j \gamma \neq 0 \right)$.
4:     Update $\beta^{(t+1)} = \beta^{(t)} + \nu \sum_{k=1}^{K} \hat{y}_k 1_{(s-1)K+k}$, where $\nu$ is the step size as shown in Algorithm 1.
5:     Compute $F^{(t)} = \rho_{\varepsilon}(\beta^{(t+1)}) + \frac{\log N}{N} \sum_{j=1}^{p} \chi \left( \hat{\beta}^{(t+1)}_j \neq 0 \right)$.
6: end for
7: Select the optimal stopping criterion by $\hat{t} = \arg\min_{1 \leq t \leq T} F^{(t)}$.
8: $\hat{\beta}^{(\hat{t})}$ is the final estimator of $\beta$. The strong learner for the combined dataset is $L_i = g \left( \sum_{j=1}^{p} \phi(X_{ij}) \hat{\beta}^{(\hat{t})}_j \right)$ with $i = 1, \ldots, N$. 
| Methods       | Commonality identification | Variable selection |
|--------------|---------------------------|-------------------|
|              | Scenario 1                |                   |
| RNP-Int      | 6.75 (0.55)               | 11.20 (0.41)      |
| NRNP-Int     | 6.70 (0.66)               | 11.90 (0.31)      |
| RP-Int       | 6.60 (0.68)               | 8.90 (0.31)       |
| NRP-Int      | 6.05 (1.00)               | 9.35 (0.59)       |
| RNP-Meta     | 1.00 (0.00)               | 11.70 (0.66)      |
| RNP-Pool     | 7.00 (0.00)               | 9.20 (0.41)       |
| RNP-Int      | 18.00 (0.00)              | 11.00 (0.00)      |
| NRNP-Int     | 17.90 (0.45)              | 9.05 (2.24)       |
| RP-Int       | 18.00 (0.00)              | 12.00 (0.00)      |
| NRP-Int      | 15.05 (2.21)              | 13.00 (1.45)      |
| RNP-Meta     | 0.55 (1.10)               | 16.30 (1.69)      |
| RNP-Pool     | 18.00 (0.00)              | 18.00 (0.00)      |
| RNP-Int      | 17.90 (0.45)              | 12.00 (0.00)      |
| NRNP-Int     | 15.85 (1.79)              | 9.05 (3.93)       |
| RP-Int       | 17.90 (0.45)              | 12.00 (0.00)      |
| NRP-Int      | 15.70 (1.63)              | 11.45 (5.37)      |
| RNP-Meta     | 1.15 (1.69)               | 15.35 (2.37)      |
| RNP-Pool     | 18.00 (0.00)              | 18.00 (0.00)      |
| RNP-Int      | 17.90 (0.45)              | 17.40 (1.23)      |
| NRNP-Int     | 13.40 (1.67)              | 3.05 (3.83)       |
| RP-Int       | 16.30 (1.65)              | 12.00 (1.26)      |
| NRP-Int      | 15.70 (2.18)              | 3.00 (3.37)       |
| RNP-Meta     | 3.20 (1.64)               | 11.70 (2.00)      |
| RNP-Pool     | 18.00 (0.00)              | 17.40 (2.00)      |
| Methods   | Commonality identification | Scenario 3 | Variable selection |
|----------|---------------------------|------------|--------------------|
|          |                           |            | TP-var  | FP-var  | RMISE   | MAE    |
| Error 1  | RNP-Int                   | 7.30 (0.66)| 2.10 (2.49) | 16.75 (2.10) | 1.10 (1.25) | 5.78 (1.49) | 3.47 (0.84) | 3.85 (0.84) | 3.47 (0.84) |
| N(0,1)   | RRNP-Int                  | 6.90 (1.37)| 0.35 (0.88) | 17.95 (0.22) | 0.85 (0.81) | 3.98 (0.99) | 2.91 (0.21) | 2.91 (0.21) | 2.91 (0.21) |
|          | RRNP-Int                  | 7.45 (0.69)| 0.70 (1.30) | 17.80 (0.52) | 1.05 (1.19) | 2.02 (1.02) | 2.75 (0.28) | 2.75 (0.28) | 2.75 (0.28) |
|          | RRNP-Int                  | 6.80 (1.06)| 0.00 (0.00) | 18.00 (0.00) | 0.65 (0.93) | 1.58 (0.38) | 2.62 (0.15) | 2.62 (0.15) | 2.62 (0.15) |
|          | RRNP-Int                  | 0.95 (0.22)| 0.45 (1.10) | 13.40 (1.05) | 3.35 (2.41) | 5.40 (2.28) | 3.23 (0.62) | 3.23 (0.62) | 3.23 (0.62) |
|          | RRNP-Int                  | 8.00 (0.00)| 19.00 (0.00) | 12.60 (1.67) | 5.10 (0.85) | 25.46 (1.42) | 12.34 (1.02) | 12.34 (1.02) | 12.34 (1.02) |
| Error 2  | RNP-Int                   | 6.30 (1.30)| 2.90 (2.10) | 15.90 (1.62) | 1.10 (0.97) | 6.78 (2.36) | 14.09 (6.35) | 14.09 (6.35) | 14.09 (6.35) |
| 70% N(0,1)| RRNP-Int                  | 6.35 (1.39)| 7.45 (5.18) | 11.30 (5.15) | 23.55 (8.49) | 54.57 (9.12) | 26.15 (6.38) | 26.15 (6.38) | 26.15 (6.38) |
|          | RRNP-Int                  | 6.65 (1.23)| 1.90 (1.48) | 17.15 (0.99) | 1.20 (1.06) | 3.39 (1.42) | 13.33 (6.35) | 13.33 (6.35) | 13.33 (6.35) |
|          | RRNP-Int                  | 6.90 (1.29)| 7.85 (5.47) | 11.00 (5.45) | 23.45 (9.31) | 55.4 (10.84) | 23.75 (6.37) | 23.75 (6.37) | 23.75 (6.37) |
|          | RRNP-Int                  | 2.10 (0.45)| 0.85 (1.04) | 16.75 (1.48) | 6.00 (2.90) | 7.37 (3.01) | 14.01 (6.37) | 14.01 (6.37) | 14.01 (6.37) |
|          | RRNP-Int                  | 8.00 (0.00)| 19.00 (0.00) | 12.55 (1.43) | 5.00 (0.73) | 25.60 (1.19) | 22.40 (6.35) | 22.40 (6.35) | 22.40 (6.35) |
| Error 3  | RNP-Int                   | 6.85 (1.27)| 5.50 (2.54) | 13.35 (2.25) | 1.90 (1.77) | 10.89 (3.02) | 15.08 (1.70) | 15.08 (1.70) | 15.08 (1.70) |
| Cauchy(0,1)| RRNP-Int                  | 6.95 (1.28)| 14.85 (2.21) | 4.05 (2.20) | 32.50 (3.80) | 245.89 (44.16) | 77.03 (11.67) | 77.03 (11.67) | 77.03 (11.67) |
|          | RRNP-Int                  | 7.05 (0.69)| 3.85 (3.00) | 15.45 (2.31) | 1.75 (1.45) | 7.92 (3.22) | 13.94 (1.79) | 13.94 (1.79) | 13.94 (1.79) |
|          | RRNP-Int                  | 7.00 (1.12)| 14.55 (2.59) | 3.90 (0.45) | 31.65 (4.33) | 269.4 (50.36) | 73.30 (11.17) | 73.30 (11.17) | 73.30 (11.17) |
|          | RRNP-Int                  | 3.75 (1.25)| 4.30 (3.26) | 11.80 (2.19) | 15.75 (3.95) | 20.58 (5.42) | 17.00 (1.76) | 17.00 (1.76) | 17.00 (1.76) |
|          | RRNP-Int                  | 8.00 (0.00)| 19.00 (0.00) | 10.25 (1.59) | 5.05 (2.04) | 27.92 (2.18) | 21.91 (1.66) | 21.91 (1.66) | 21.91 (1.66) |
**TABLE 3** Data analysis: number of overlapping genes. In each cell, male/female for GBM data, and stage I/II/III and IV for LUAD data

|                | RNP-Int | NRNP-Int | RP-Int | NRP-Int | RNP-Meta | RNP-Pool |
|----------------|---------|----------|--------|---------|----------|----------|
| **GBM**        |         |          |        |         |          |          |
| RNP-Int        | 5/11    | 3/5      | 2/2    | 2/3     | 5/3      |          |
| NRNP-Int       | 9/9     | 2/4      | 2/3    | 3/3     | 3/1      |          |
| RNP-Meta       | 5/9     | 4/6      | 0/2    | 2/0     |          |          |
| RNP-Pool       | 6/11    | 0/2      | 2/0    |         |          |          |
| RP-Int         |         |          |        |         |          | 8/10     |
| NRP-Int        |         |          |        |         |          | 2/0      |
| **LUAD**       |         |          |        |         |          | 5/5      |
| RNP-Int        | 14/15/16| 8/7/6    | 2/3/3  | 1/1/3   | 3/3/7    | 14/15/15 |
| NRNP-Int       | 18/13/14| 1/1/1    | 1/0/0  | 3/3/5   | 8/7/6    |          |
| RP-Int         | 10/10/11| 3/2/4    | 1/3/1  | 3/3/3   |          |          |
| NRP-Int        |         | 14/13/17 | 2/3/2  | 2/1/3   |          |          |
| RNP-Meta       |         |          |        |         | 17/17/16 | 3/3/6    |
| RNP-Pool       |         |          |        |         | 16/16/16 |          |

**TABLE 4** Data analysis: number of identified genes, heterogeneities, and commonalities.

|                | Genes | Heterogeneities | Commonalities |
|----------------|-------|-----------------|---------------|
| **GBM**        |       |                 |               |
| RNP-Int        | 13    | 10              | 3             |
| NRNP-Int       | 16    | 15              | 1             |
| RP-Int         | 12    | 10              | 2             |
| NRP-Int        | 15    | 14              | 1             |
| RNP-Meta       | 18    | 18              | 0             |
| RNP-Pool       | 5     | 0               | 5             |
| **LUAD**       |       |                 |               |
| RNP-Int        | 17    | 5               | 12            |
| NRNP-Int       | 18    | 6               | 12            |
| RP-Int         | 11    | 1               | 10            |
| NRP-Int        | 17    | 4               | 13            |
| RNP-Meta       | 40    | 40              | 0             |
| RNP-Pool       | 16    | 0               | 16            |
FIGURE C1 Illustration of different loss function. First row: loss functions $\rho(x)$. Second row: influence functions $\psi(x)$.

C MORE FIGURES AND TABLES
FIGURE C2 Simulation under Scenario 1 and error 3: estimation of the effects of the first six covariates. Three rows correspond to three datasets.

FIGURE C3 Data analysis: Average of MAE in five-fold CV as a function of $\lambda$. Left: GBM. Right: LUAD.
FIGURE C4 Analysis of GBM data using the alternative NRNP-Int method: estimated effects of 16 genes. Blue solid lines correspond to females, and orange dotted lines to males.
FIGURE C5 Analysis of GBM data using the alternative RP-Int method: estimated effects of 12 genes. Blue solid lines correspond to females, and orange dotted lines to males.
FIGURE C6 Analysis of GBM data using the alternative NRP-Int method: estimated effects of 15 genes. Blue solid lines correspond to females, and orange dotted lines to males.
FIGURE C7 Analysis of GBM data using the alternative RNP-Meta method: estimated effects of 18 genes. Blue solid lines correspond to females, and orange dotted lines to males.

FIGURE C8 Analysis of GBM data using the alternative RNP-Pool method: estimated effects of 5 genes. Blue solid lines correspond to females, and orange dotted lines to males.
FIGURE C9 Analysis of TCGA LUAD data. Left: scatter plot of FEV1. Right: FEV1 against ARHGAP4 gene expression, and regression curve with bootstrap-based 95% confidence intervals (shaded area).
FIGURE C10 Analysis of LUAD data using the proposed RNP-Int method: estimated effects of 17 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
FIGURE C11 Analysis of LUAD data using the proposed NRNP-Int method: estimated effects of 18 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
FIGURE C12. Analysis of LUAD data using the proposed RP-Int method: estimated effects of 11 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
FIGURE C13 Analysis of LUAD data using the proposed NRP-Int method: estimated effects of 17 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
FIGURE C14 Analysis of LUAD data using the proposed RNP-Meta method: estimated effects of 40 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
FIGURE C15 Analysis of LUAD data using the proposed RNP-Pool method: estimated effects of 16 genes. Purple lines correspond to stage I, green lines to stage II, and yellow lines to stage III and IV.
## TABLE C1 Simulation Scenarios 1-2 for data with a right-censored survival outcome. In each cell, mean (standard deviation).

| Commonality identification | Variable selection | Scenario 1 | Scenario 2 |
|---------------------------|-------------------|------------|------------|
|                           | Methods           | TP-ind     | FP-ind     | TP-var | FP-var | RMISE | Cstat |
|                           | RNP-Int           | 6.10 (1.17) | 1.75 (0.79) | 11.10 (0.45) | 0.30 (0.47) | 2.58 (0.81) | 0.90 (0.01) |
| Error 1                   | NRNP-Int          | 6.30 (0.87) | 0.50 (0.89) | 11.80 (0.41) | 1.65 (1.18) | 2.57 (0.42) | 0.90 (0.01) |
| N(0,1)                    | RP-Int            | 6.30 (1.03) | 4.30 (0.98) | 8.80 (0.62) | 0.40 (0.60) | 4.38 (1.17) | 0.85 (0.03) |
|                           | NRP-Int           | 5.45 (1.40) | 3.45 (0.69) | 9.30 (0.57) | 13.10 (3.16) | 6.21 (1.20) | 0.82 (0.03) |
|                           | RNP-Meta          | 1.00 (0.32) | 0.85 (1.27) | 11.20 (1.11) | 6.20 (2.35) | 5.62 (2.42) | 0.87 (0.04) |
|                           | RNP-Pool          | 7.00 (0.00) | 11.00 (0.00) | 9.25 (0.55) | 3.35 (0.75) | 10.80 (0.81) | 0.71 (0.03) |
| Error 2                   | RNP-Int           | 6.45 (0.95) | 2.05 (0.83) | 11.00 (0.32) | 0.5 (0.69) | 2.62 (0.98) | 0.86 (0.03) |
| 70% N(0,1)                | NRNP-Int          | 5.45 (1.23) | 5.50 (3.36) | 6.85 (4.08) | 18.75 (5.16) | 48.37 (9.08) | 0.66 (0.11) |
| +30% Cauchy(0,1)          | RP-Int            | 6.05 (0.89) | 4.05 (0.39) | 8.90 (0.45) | 0.50 (0.76) | 4.23 (0.79) | 0.82 (0.03) |
|                           | NRP-Int           | 5.70 (1.22) | 6.80 (3.00) | 5.30 (3.85) | 20.85 (3.96) | 56.16 (10.52) | 0.63 (0.11) |
|                           | RNP-Meta          | 1.00 (0.32) | 1.95 (1.40) | 10.25 (1.25) | 9.50 (2.67) | 8.58 (3.37) | 0.81 (0.05) |
|                           | RNP-Pool          | 7.00 (0.00) | 11.00 (0.00) | 9.15 (0.49) | 3.30 (0.80) | 10.83 (6.01) | 0.69 (0.03) |
| Error 3                   | RNP-Int           | 6.00 (1.17) | 3.30 (2.20) | 9.70 (2.45) | 2.10 (2.13) | 6.42 (4.79) | 0.76 (0.05) |
| Cauchy(0,1)               | NRNP-Int          | 6.45 (1.00) | 9.30 (2.03) | 1.70 (2.43) | 22.15 (3.00) | 237.62 (45.4) | 0.49 (0.07) |
|                           | RP-Int            | 6.30 (0.80) | 4.55 (1.05) | 8.35 (1.14) | 0.85 (1.04) | 5.39 (1.76) | 0.85 (0.04) |
|                           | NRP-Int           | 5.95 (1.47) | 9.85 (1.46) | 1.00 (1.56) | 22.65 (3.41) | 16.82 (4.37) | 0.69 (0.05) |
|                           | RNP-Meta          | 1.85 (0.88) | 4.20 (1.58) | 7.15 (1.63) | 16.50 (3.41) | 14.20 (1.00) | 0.64 (0.04) |
|                           | RNP-Pool          | 7.00 (0.00) | 11.00 (0.00) | 8.25 (2.34) | 4.80 (2.80) | 14.20 (1.00) | 0.64 (0.04) |
| Methods          | Commonality identification | Variable selection | Scenario 3 |
|------------------|---------------------------|--------------------|------------|
|                  | TP-ind | FP-ind | TP-var | FP-var | RMISE | Cstat |
| **Error 1**      |         |        |        |        |       |       |
| N(0,1)           | RNP-Int | 0.00 (0.00) | 3.50 (2.35) | 12.00 (1.92) | 0.95 (1.19) | 7.40 (3.41) | 0.90 (0.03) |
| RNP-Int          | 0.00 (0.00) | 1.50 (0.95) | 13.35 (0.67) | 2.10 (1.94) | 5.66 (1.43) | 9.22 (2.24) | 0.86 (0.02) |
| RP-Int           | 0.00 (0.00) | 7.65 (1.73) | 9.55 (1.00) | 1.50 (1.40) | 9.22 (2.24) | 7.62 (2.58) | 0.83 (0.02) |
| NRP-Int          | 0.00 (0.00) | 6.65 (1.46) | 10.00 (0.65) | 12.45 (3.47) | 15.88 (2.76) | 18.45 (2.74) | 26.22 (6.00) |
| RNP-Meta         | 0.00 (0.00) | 0.45 (1.10) | 12.25 (1.52) | 9.20 (4.12) | 9.82 (5.22) | 6.60 (2.06) | 0.63 (0.05) |
| RNP-Pool         | 0.00 (0.00) | 18.00 (0.00) | 11.25 (2.40) | 2.85 (0.99) | 19.40 (1.39) | 31.19 (4.82) | 0.72 (0.04) |
| **Error 2**      |         |        |        |        |       |       |
| 70% N(0,1)       | RNP-Int | 0.00 (0.00) | 0.46 (2.26) | 11.90 (1.17) | 1.30 (1.59) | 8.41 (3.80) | 0.83 (0.03) |
| +30% Cauchy(0,1) | NRP-Int | 0.00 (0.00) | 9.35 (4.58) | 7.75 (3.61) | 16.70 (5.66) | 52.78 (9.62) | 0.71 (0.11) |
| RNP-Int          | 0.00 (0.00) | 7.65 (1.73) | 9.55 (1.00) | 1.50 (1.40) | 9.22 (2.24) | 7.62 (2.58) | 0.83 (0.02) |
| NRP-Int          | 0.00 (0.00) | 6.65 (1.46) | 10.00 (0.65) | 12.45 (3.47) | 15.88 (2.76) | 18.45 (2.74) | 26.22 (6.00) |
| RNP-Meta         | 0.00 (0.00) | 0.45 (1.10) | 12.25 (1.52) | 9.20 (4.12) | 9.82 (5.22) | 6.60 (2.06) | 0.63 (0.05) |
| RNP-Pool         | 0.00 (0.00) | 18.00 (0.00) | 11.25 (2.40) | 2.85 (0.99) | 19.40 (1.39) | 31.19 (4.82) | 0.72 (0.04) |
| **Error 3**      |         |        |        |        |       |       |
| Cauchy(0,1)      | RNP-Int | 0.00 (0.00) | 0.46 (2.26) | 11.90 (1.17) | 1.30 (1.59) | 8.41 (3.80) | 0.83 (0.03) |
| RNP-Int          | 0.00 (0.00) | 15.40 (2.98) | 2.30 (2.54) | 22.00 (3.34) | 247.11 (48.35) | 0.51 (0.09) |
| NRP-Int          | 0.00 (0.00) | 15.40 (2.98) | 2.30 (2.54) | 22.00 (3.34) | 247.11 (48.35) | 0.51 (0.09) |
| RNP-Meta         | 0.00 (0.00) | 15.40 (2.98) | 2.30 (2.54) | 22.00 (3.34) | 247.11 (48.35) | 0.51 (0.09) |
| RNP-Pool         | 0.00 (0.00) | 18.00 (0.00) | 10.50 (1.57) | 4.20 (2.67) | 20.61 (3.03) | 0.67 (0.03) |

**TABLE C2 Simulation Scenarios 3-4 for data with a right-censored survival outcome. In each cell, mean (standard deviation).**
**TABLE C3** Simulation Scenarios 1 for data with a continuous outcome under unbalanced subgroups. In each cell, mean (standard deviation).

| Commonality identification | Variable selection |
|---------------------------|--------------------|
| Methods                   | TP-ind             | FP-ind | TP-var | FP-var | RMISE  | MAE    |
| RNP-Int                   | 0.00 (0.00)        | 3.50 (2.35) | 12.00 (1.92) | 0.95 (1.19) | 7.40 (3.41) | 4.42 (0.81) |
| NRNP-Int                  | 0.00 (0.00)        | 1.50 (0.95) | 13.35 (0.67) | 2.10 (1.94) | 5.66 (1.43) | 3.80 (0.55) |
| TP-var                    | 0.95 (1.00)        | 1.50 (1.40) | 9.22 (2.24)  | 5.98 (0.85)  |           |        |
| RP-Int                    | 6.65 (1.46)        | 12.45 (3.47) | 15.88 (2.76) | 7.05 (0.98)  |           |        |
| FP-var                    | 9.20 (4.12)        | 5.98 (0.85)  | 12.20 (1.19) | 6.90 (0.32)  |           |        |
| NRP-Int                   | 0.45 (1.10)        | 12.25 (1.52) | 9.82 (5.22)  | 5.08 (1.88)  |           |        |
| RNP-Meta                  | 18.00 (0.00)       | 2.85 (0.99)  | 19.40 (1.39) | 10.92 (1.91) |           |        |

**TABLE C4** Simulation Scenarios 1 for data with a right-censored survival outcome under unbalanced subgroups. In each cell, mean (standard deviation).

| Commonality identification | Variable selection |
|---------------------------|--------------------|
| Methods                   | TP-ind             | FP-ind | TP-var | FP-var | RMISE  | Cstat  |
| RNP-Int                   | 5.80 (1.36)        | 2.05 (1.15) | 10.95 (0.61) | 0.9 (1.33) | 3.40 (0.98) | 0.89 (0.02) |
| NRNP-Int                  | 6.25 (1.12)        | 0.35 (0.75) | 11.80 (0.41) | 2.21 (1.32) | 3.03 (1.20) | 0.90 (0.01) |
| RP-Int                    | 6.40 (0.75)        | 4.40 (0.82) | 8.60 (0.82)  | 1.05 (1.15) | 4.78 (1.59) | 0.85 (0.02) |
| NRP-Int                   | 5.55 (1.23)        | 3.80 (0.77) | 9.35 (0.99)  | 11.50 (3.53) | 9.21 (2.01) | 0.85 (0.02) |
| RNP-Meta                  | 1.00 (0.32)        | 1.60 (1.19) | 9.65 (1.09)  | 8.75 (2.17) | 8.81 (2.44) | 0.86 (0.02) |
| RNP-Pool                  | 7.00 (0.00)        | 11.00 (0.00) | 9.40 (0.68)  | 3.65 (1.14) | 10.80 (0.61) | 0.75 (0.03) |

| Error 2                   | RNP-Int             | 6.20 (1.20) | 2.50 (1.24) | 10.60 (0.88) | 1.10 (1.29) | 4.00 (2.41) | 0.85 (0.03) |
| 70% N(0,1)                | NRNP-Int            | 6.57 (2.11) | 5.35 (3.17) | 6.70 (3.54)  | 18.10 (4.93) | 46.21 (8.96) | 0.65 (0.13) |
| +30%Cauchy(0,1)           | RP-Int             | 6.10 (0.85) | 4.55 (1.19) | 8.60 (0.82)  | 1.20 (1.06) | 5.22 (1.59) | 0.82 (0.02) |
|                           | NRP-Int             | 6.25 (0.85) | 6.80 (2.59) | 5.25 (3.28)  | 19.85 (3.23) | 54.40 (10.72) | 0.64 (0.13) |
|                           | RNP-Meta            | 1.10 (0.45) | 2.80 (1.70) | 8.90 (1.68)  | 11.60 (2.80) | 11.60 (3.38) | 0.80 (0.04) |
|                           | RNP-Pool            | 7.00 (0.00) | 11.00 (0.00) | 9.05 (0.22)  | 3.40 (1.00) | 11.00 (1.00) | 0.73 (0.03) |

| Error 3                   | RNP-Int             | 6.25 (1.45) | 3.65 (1.42) | 9.40 (1.96)  | 2.70 (1.72) | 7.60 (3.40) | 0.77 (0.03) |
| Cauchy(0,1)               | NRNP-Int            | 6.40 (0.75) | 9.20 (1.96) | 1.75 (2.36)  | 22.40 (2.50) | 232.22 (40.43) | 0.50 (0.08) |
|                           | RP-Int             | 6.30 (0.98) | 4.65 (0.99) | 8.25 (1.18)  | 1.75 (1.37) | 7.40 (1.82) | 0.75 (0.03) |
|                           | NRP-Int             | 6.10 (0.97) | 9.70 (1.46) | 1.40 (1.67)  | 22.45 (3.50) | 300.78 (55.32) | 0.50 (0.09) |
|                           | RNP-Meta            | 1.45 (0.69) | 3.80 (1.28) | 7.35 (1.53)  | 16.05 (3.02) | 17.89 (3.22) | 0.71 (0.04) |
|                           | RNP-Pool            | 7.00 (0.00) | 11.00 (0.00) | 8.95 (1.36)  | 3.50 (0.89) | 12.20 (1.19) | 0.69 (0.03) |