High Dimensional Feature Selection algorithms with Interactions on Time-to-Event Outcome

Lin Yu

Supervisor: Dr. Wei Xu

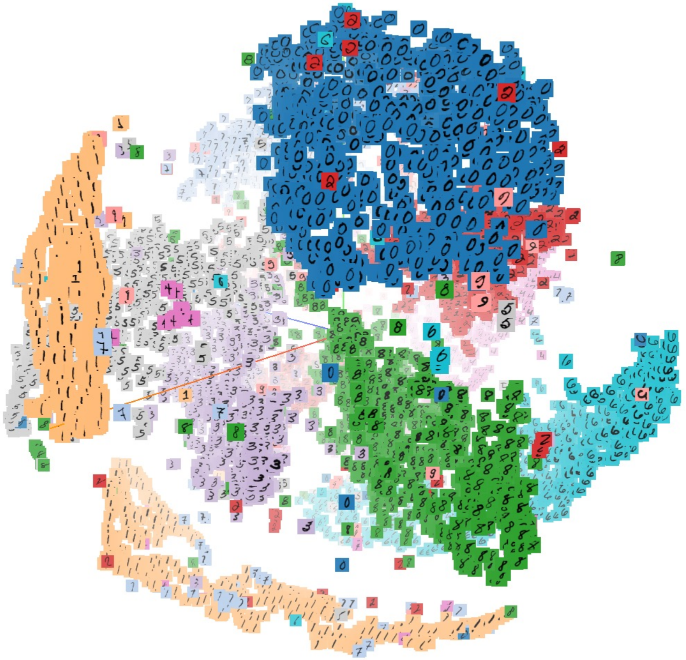
Biostatistics Department, University Health Network

2024-04-25

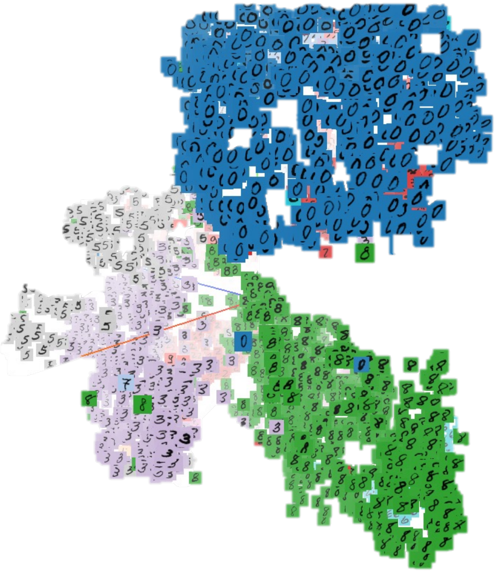
UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

UHN Princess Margaret Cancer Centre

Background: Challenges of High Dimensional Data



Feature selection (e.g., LASSO, Ridge)

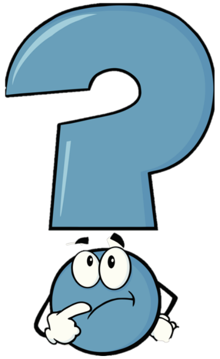


High dimensional data:
model overfitting, generalizability

Low dimensional data

Image copyright: Visualize high dimensional data. (pinterest.com)

Research Question and Objective



Interactive effects not considered

Existing methods:

HDSI algorithms[1,2,3] for continuous and binary cases

Research question: Is HDSI/RHDSI Robust? Can it be extended to different types of data?

Objective: Develop feature selection algorithm with **interactions** for **time-to-event outcome**

1. Jain R, Xu W. HDSI: High dimensional selection with interactions algorithm on feature selection and testing. PLOS ONE.
2. Jain R, Xu W. RHDSI: A novel dimensionality reduction based algorithm on high dimensional feature selection with interactions. Inf Sci. 2021 Oct 1;574:590–605.
3. Zhuang Z, Xu W, Jain R. High Dimensional Selection with Interactions Algorithm on Feature Selection for Binary Outcome.

Method Pipeline

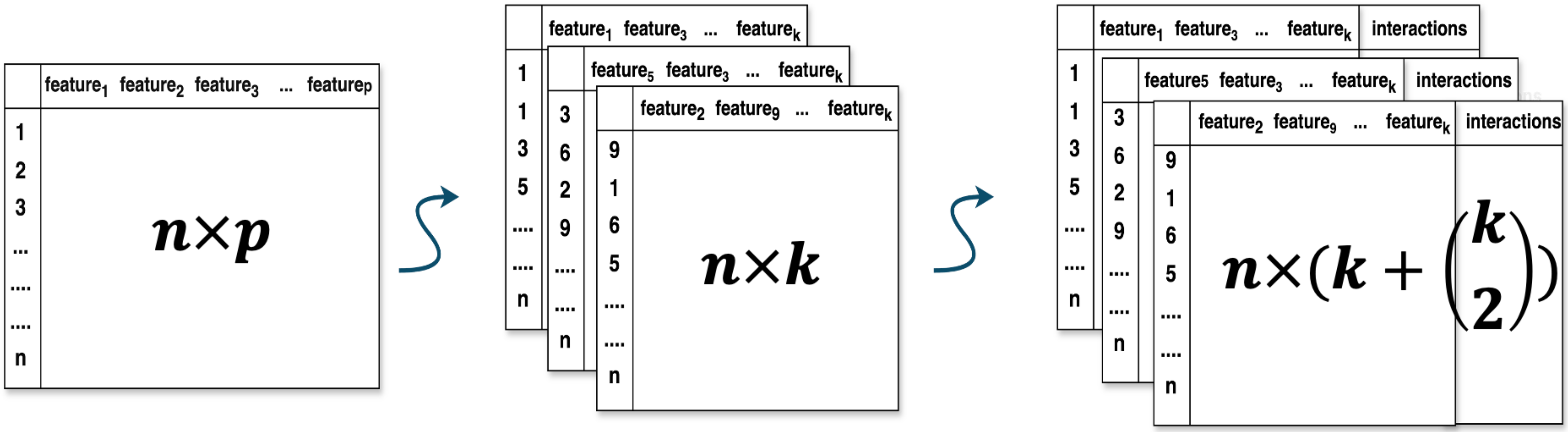
1: Develop algorithms for model building and hyper parameters tuning

2: Conduct simulations with high dimensional features with both marginal and interactive effects

3: Implement the proposed algorithms into real clinical study

Method: Development of the HDSI-LASSO and HDSI-Ridge Algorithms

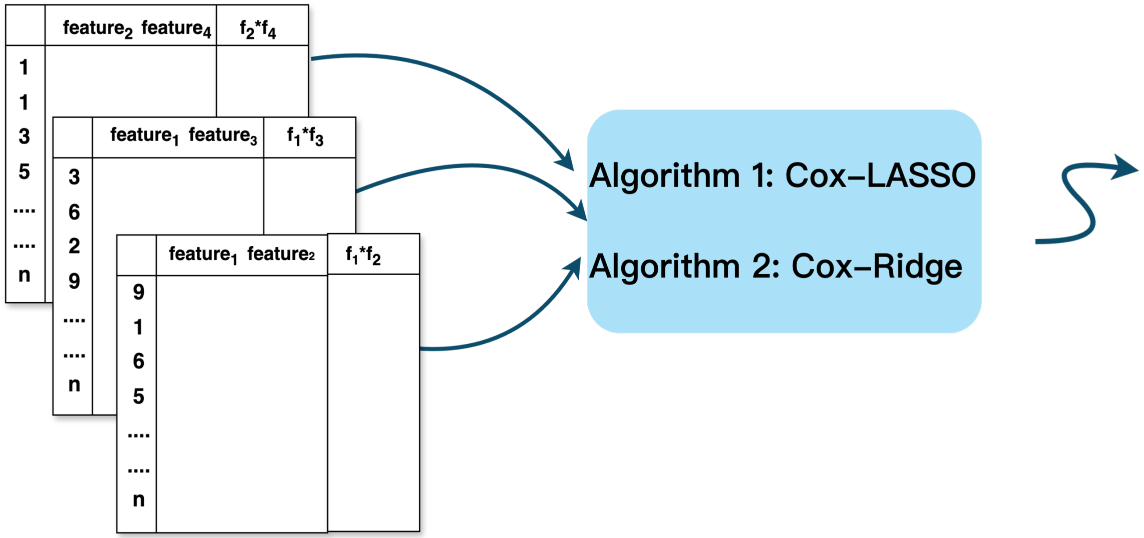
Step 1: Prepare Bootstrap Sets



B Bootstrap datasets with interactions

Method: Development of the HDSI-LASSO and HDSI-Ridge Algorithms

Step 2: Build model and select features



Input

Feature Selection Algorithm

bootstrap set 1 output		
X_1	X_2	X_{1*2}
$\hat{\beta}_1^1$	$\hat{\beta}_2^1$	$\hat{\beta}_{1*2}^1$
C_1^1	C_2^1	C_{1*2}^1

...

bootstrap set B output		
X_2	X_4	X_{2*4}
$\hat{\beta}_2^3$	$\hat{\beta}_4^3$	$\hat{\beta}_{2*4}^3$
C_2^3	C_4^3	C_{2*4}^3

Output:
- **Coef. estimates($\hat{\beta}$)**
- **C-index(C)**

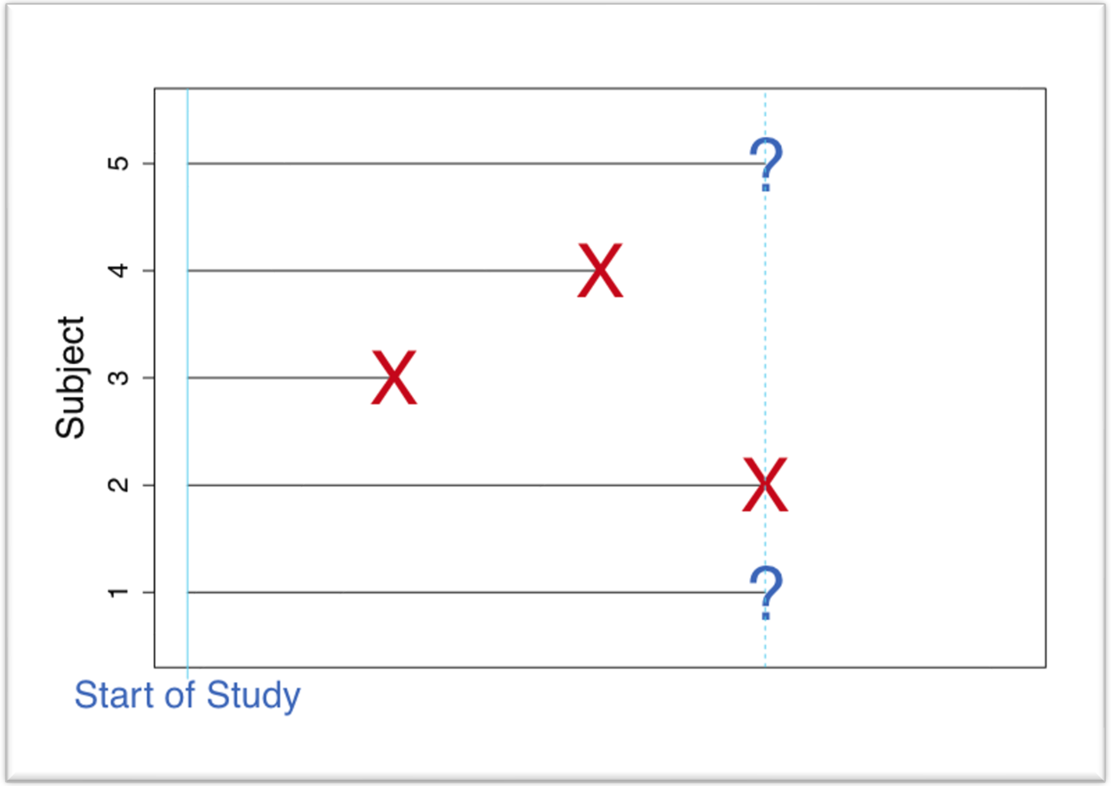
Pooled results:

- $\hat{\beta}_j = \text{avg}(\hat{\beta}_j^1, \hat{\beta}_j^2, \dots \hat{\beta}_j^B)$
- $\text{Cindex}(X_j) = \min(C_j^1, C_j^2, \dots C_j^B)$

Criteria for feature selection

- Significance of X_j (quantile includes 0)
- $\text{Cindex}(X_j) > \text{cutoff value}$

Simulation Study Design



Copyright: taken from Prof. Kevin E. Thorpe's lecture slides

1) Observed time T

$$T = \min(\tilde{T}, C)$$

$\tilde{T}$: The latent time had everyone's survival time observed

C: The censoring time

2) Observed status Y

$$Y = \begin{cases} \text{event}, & T = \tilde{T} \\ \text{censored}, & T = C \end{cases}$$

Simulation Study Design

Step 1) Latent event time $\tilde{T}$

- Survival function

$$S(\tilde{T}) = 1 - F(\tilde{T}) \sim \text{Unif}\,(0,1)$$

- Cox model

$$S(\tilde{T}|x) = \exp[-H_0(\tilde{T})\exp(Z)]$$

H_0: cumulative baseline hazard

Z: linear predictor

- Inverse of survival function

$$\boldsymbol{\tilde{T}} = \boldsymbol{H_0}^{-1}(-\boldsymbol{log}(\boldsymbol{S})\exp(-\boldsymbol{Z}))$$

Step 2) Censoring time C

$$C \sim \text{Unif}\,(0, b)$$

Step 3) Compare $\tilde{T}$ and C

$$T = \min(\tilde{T}, C)$$

$$Y = \begin{cases} \text{event}, & T = \tilde{T} \\ \text{censored}, & T = C \end{cases}$$

Simulation Study Design

True model

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(Z)$$

$$= \lambda_0(t)\exp(X_1 + X_2 + 0.75X_3 - 0.75X_4 + 0.75X_5 + X_1X_2 - X_3X_4)$$

$X_1, X_2, \ldots X_5$: continuous, generated from multinormal

Samples

1000 for training; 500 for testing (Event rate: ~ 40%)

Features

True

- 5 marginal + 2 interactive

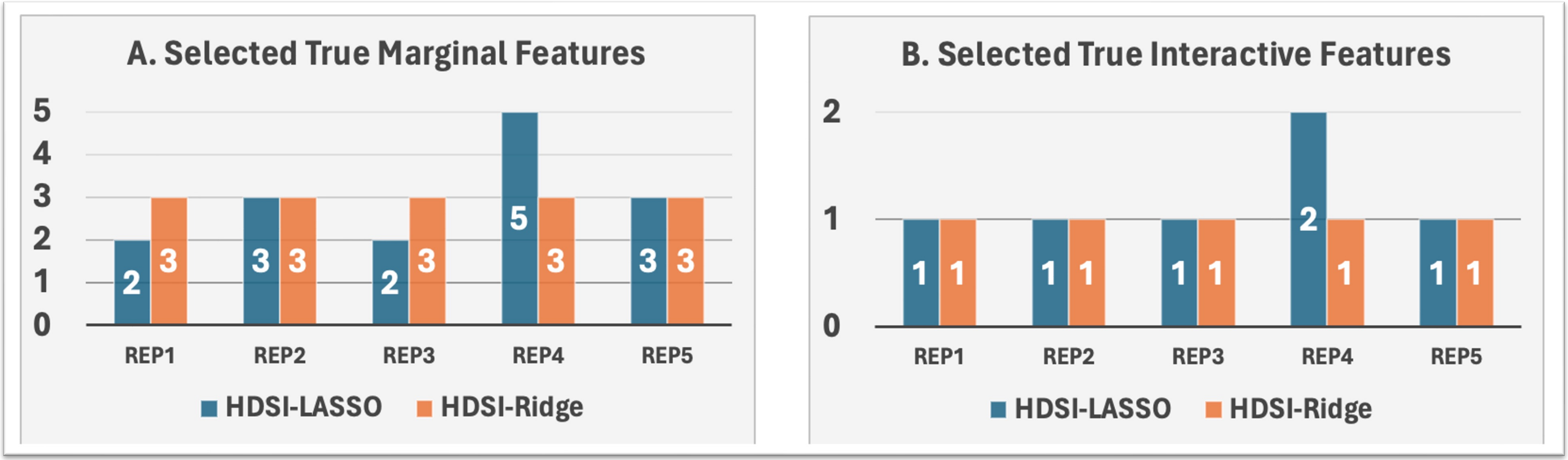
Noisy

- 20 marginal + 298 interactive

Simulation Study Results

True model: 5 true marginal, 2 true interactive features

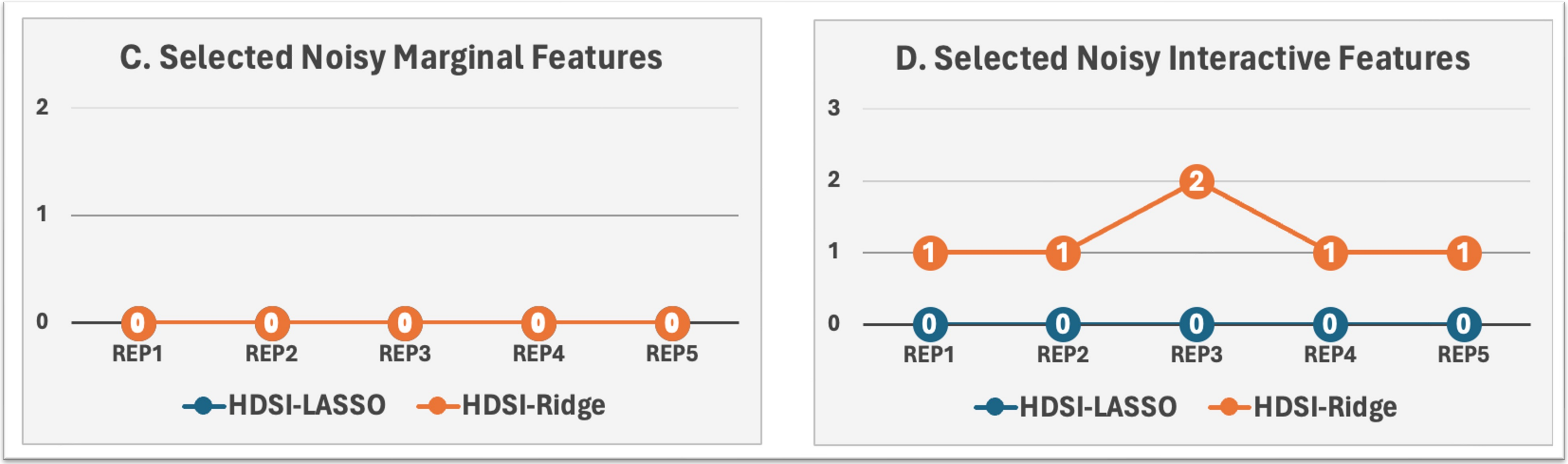
Are all true effective features selected?



Simulation Study Results

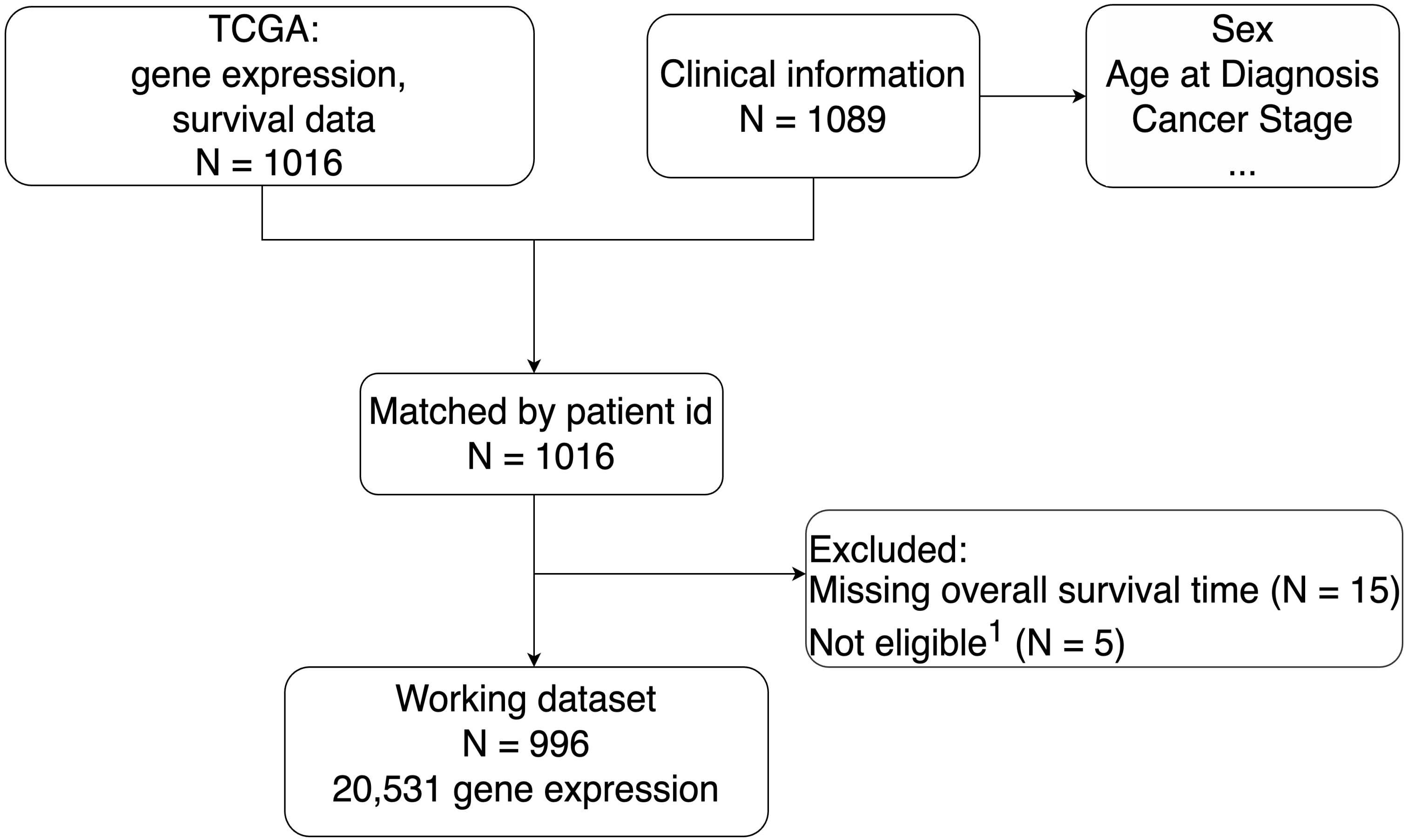
Noisy: 20 noisy marginal, 298 noisy interactive features

Are any noisy features selected?



Real-World Study

Setting: relationship between gene expression profile & overall survival in lung cancer patients?



Summary statistics:

Event rate: 40%

Median survival time: ~2 yrs

Univariate analysis:

- Age at diagnosis
- cancer stage
- Top 50 and 100 significant genes

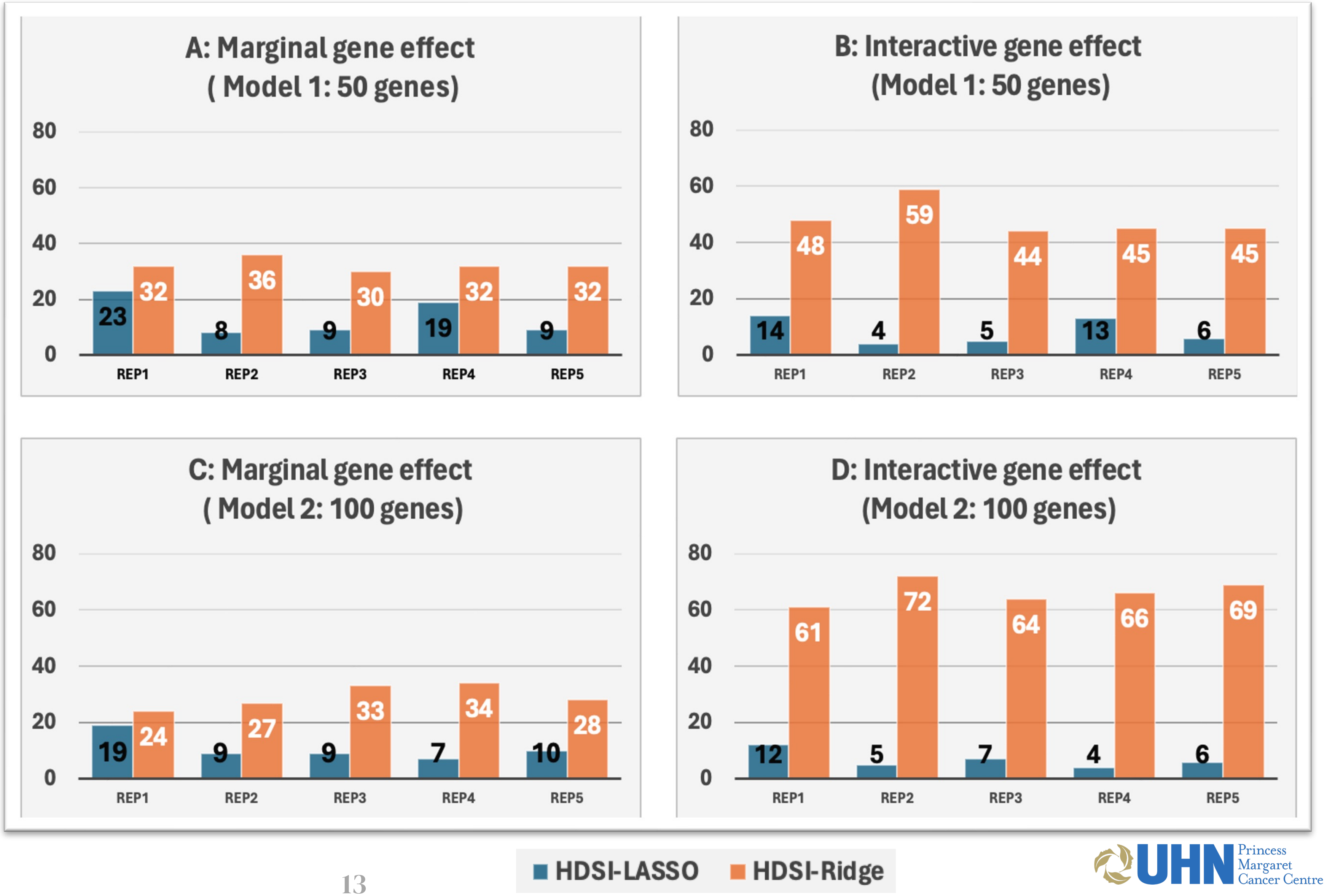
1: subjects who died or censored at the enrollment

Real-World Study Results

Model 1 : 50 marginal + 1225 interactive; **Model 2**: 100 marginal + 4950 interactive

Summary:

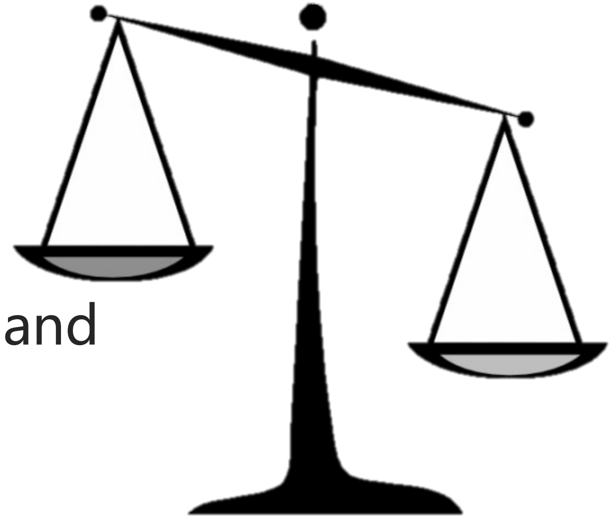
- HDSI-Ridge selected more genes
- Marginal features: Both robust
- Interactive features: only HDSI-LASSO robust
- C-index: HDSI-Ridge> HDSI-LASSO



Discussion

HDSI-LASSO:

- Selected **less** features
- **Robust** to the increase in the number of features(marginal and interactive)
- Slightly lower C-index



HDSI-Ridge:

- Selected more features
- Only robust to the increase in the number of marginal features; **Selected more noisy features**

Limitations and future work:

- Corporate other algorithms into the HDSI framework
- Consider other simulation settings (e.g., different effect sizes)

References

1. Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723. https://doi.org/10.1002/sim.2059

2. Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K. E., McLellan, M., Sofia, H. J., Hutter, C., Getz, G., Wheeler, D., Ding, L., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., … Mariamidze, A. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems*, *6*(3), 271-281.e7. https://doi.org/10.1016/j.cels.2018.03.002

3. Jain, R., & Xu, W. (n.d.). HDSI: High dimensional selection with interactions algorithm on feature selection and testing. *PLOS ONE*.

4. *The Cancer Genome Atlas Program (TCGA)—NCI* (nciglobal,ncienterprise). (2022, May 13). [cgvMiniLanding]. https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Theng, D., & Bhoyar, K. K. (2023). Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-023-02010-5

5. Zhuang, Z., Xu, W., & Jain, R. (n.d.). *High Dimensional Selection with Interactions Algorithm on Feature Selection for Binary Outcome*.

Thank you ☺