January 2021

Sex Difference In Identification Of Predictive Tumor Tissue Metabolites Associated With Colorectal Cancer Prognosis

Xinyi Shen
xinyi.shen@yale.edu

Yale SCHOOL OF PUBLIC HEALTH

Sex Difference in Identification of Predictive Tumor Tissue Metabolites Associated with Colorectal Cancer Prognosis

Xinyi Shen

Year Completed: 2021

Year Degree Awarded: 2021

Degree Awarded: Master of Public Health

Department of Chronic Disease Epidemiology, Yale School of Public Health

Advisor: Dr. Caroline Johnson

Committee Member: Dr. Harvey Risch

Abstract

Colorectal cancer (CRC) is the third major cause of cancer-related deaths in the United States in 2020. Sex-related differences in CRC stage, prognosis, and metabolism have become increasingly popular in cancer research. Males have poorer survival for CRC, but females with right-sided colon cancer (RCC) have aberrant metabolism correlated with poor survival. Delay in knowing the condition of CRC in female patients would result in poor prognosis, which could be avoided by predicting prognostic outcomes. Random Survival Forest (RSF) is ideal for exploration and making predictions using metabolomics data with high dimension, strong collinearity, and heterogeneity, which CPH models could not efficiently address. In this retrospective study including 197 patients, we applied an RSF prediction method based on the backward selection algorithm in 5-year overall survival (OS) for 95 female CRC patients and validated its performance. We also investigated Cox proportional hazard models (CPH), lasso penalized Cox regression (Cox-Lasso), and Logistic Regression (LR) and compared their predictive performances. RSF using the backward selection algorithm showed the best performance with the C-index of the training and testing sets reaching 0.81(95% CI: 0.810-0.813) and 0.78 (95% CI: 0.776-0.777) respectively and identified the five most predictive metabolites for female 5-year OS: glutathione, citrulline, phosphoenolpyruvate, lysoPC (16:0), and asparagine. Accordingly, the backward selection algorithm-based Random Survival Forest model using tumor tissue metabolic profile is promising for predicting 5-year OS for female CRC patients. The results could be easily interpreted and applied in preventive medicine and precision medicine, guiding clinicians in choosing targeted treatments by sex for better survival and avoiding unnecessary treatments.

Acknowledgment

I would like to express my deepest gratitude to Dr. Caroline Johnson and Dr. Harvey Risch, my research supervisors, for their patient guidance, enthusiastic encouragement, and useful critiques of this research work. I would also like to thank Dr. Sajid Khan, for his valuable advice and assistance in results interpretation from an oncologist's perspective. My grateful thanks are also extended to Dr. Lingeng Lu and Dr. Yawei Zhang for their advice on epidemiology methodology and data analysis, to Dr. Hong Yan, who offered me great help in addressing overfitting problems and to, Dr. Yuping Cai and Dr. Huang Huang for their support during the critical period for the introduction of learning metabolomics and survival analysis. I would also like to extend my thanks to Dr. Andrew DeWan, my academic advisor, for his recommendation of Johnson laboratory and academic support during the two years.

Besides, I am particularly grateful for the spiritual support throughout my hardship from my close friends at Yale who have been so generous and encouraging.

Finally, I wish to thank my parents for their support and encouragement throughout my study both financially and mentally.

Contents

List of Tables **1**
List of Figures **5**
Introduction **13**
Methods and materials **16**
Sample Collection and Metabolites Measurements 16
Statistical Analysis 16
Random survival forest (RSF) 17
LASSO-based CPH 17
Logistic Regression (LR) 17
Modeling process 18
Results **19**
Clinical characteristics 19
Cox proportional hazard regression (CPH) analysis 19
Predictive modeling 21
Primary exploration of the possibility to predict CRC prognosis using tumor tissue untargeted metabolic profile *21*
Cox proportional hazard regression (CPH) models with variable selection using forward stepwise and LASSO *23*
Random survival forest (RSF) model *24*
Logistic regression model (LR) *25*
Comparing predictive performance for the models *26*
Discussion **28**
Conclusion **34**
References **36**

Appendix is attached at the end.

List of Tables

Table 1. Demographic Characteristics and Clinical Factors

Characteristics	No. of Patients	5-Year Overall Survival (OS)			5-Year Recurrence-free Survival (RFS)		
		Deaths, No.	Rate, %[a]	*P*[*]	Cases, No.	Rate, %[a]	*P*[*]
Age at diagnosis, y							
55-60	19	2	88.9		3	84.2	
61-69	64	9	83.4	0.048	14	72.4	0.208
70-79	81	15	78.7		12	81.4	
≥80	33	11	61.5		1	96.3	
Sex, n							
Male	102	23	74.3	0.176	17	77.5	0.478
Female	95	14	83.2		13	84.0	
Chemotherapy, n							
Yes	66	18	68.5	0.026	15	73.1	0.027
No	131	19	83.7		15	85.0	
Clinical stage, n							
I	47	3	92.5		5	88.4	
II	86	13	82.4	0.001	11	82.0	0.091
III	64	21	63.5		14	73.8	
Anatomic tumor location, n							
Left	99	17	81.2	0.422	19	77.2	0.230
Right	98	20	75.6		11	85.3	

[*] *P* value of Log-rank test.

[a] Survival rates were calculated by the Kaplan-Meier estimation method.

Table 2. Results of fitting Cox model including sex, anatomic tumor location, clinical stage, and age on 5-year OS and RFS of patients with colorectal cancer (n = 197)

Variables	OS			RFS		
	HR	95% CI	P	HR	95% CI	P
Sex: male	2.05	1.04-4.05	0.039	1.25	(0.60, 2.59)	0.555
Anatomic tumor location: RCC	0.84	0.42-1.68	0.623	0.72	(0.33, 1.56)	0.410
Clinical stage: Late	4.17	2.11-8.24	<0.001	2.08	(1.01, 4.27)	0.047
Age	1.09	1.05-1.14	<0.001	0.97	(0.92, 1.02)	0.261

Table 3. **Primary exploration of predictive performance using RSF for CRC prognosis. The MPER might slightly fluctuate due to randomness.** MPER: Minimum prediction error rate.

Outcome	Population	MPER	Selected metabolites
OS	All patients (n = 197)	0.3085	CDP-Ethanolamine, CMP, Dimethylsphingosine, Glutathione, disulfide, Hypoxanthine, Tyrosine
	Females (n = 95)	**0.2100**	Asparagine, Citrulline, Creatinine, DHAPorG3P, Glutathione
	Males (n = 102)	0.3407	Tyrosine, Uracil
RFS	All patients (n = 197)	0.3001	Acetyl-lysine, Cytidine, Dimethylsphingosine, Glutathione disulfide, LysoPE (22:5), Palmitic acid, Xanthosine
	Females (n = 95)	0.3257	Glutathione, Glutathione disulfide
	Males (n = 102)	**0.2356**	Acetyl-lysine, Hypoxanthine, N1-Acetylspermine, Xanthosine

Table 4. **Characteristics of training and testing sets for CRC female patients (n = 95).**

Variables	Training set (n = 58)	Testing set (n = 37)	*P**
Age (mean (SD))	70.98 (8.03)	72.11 (7.51)	0.490
Chemotherapy, n			
Yes	21	14	0.342
No	37	23	
Clinical stage			
Early	38	22	0.705
Late	20	15	
Anatomic tumor location			
Left	28	17	0.991
Right	30	20	
Death			
Yes	9	5	1.000
No	49	32	
Follow-up months (mean (SD))	48.43 (19.30)	44.79 (21.87)	0.560

* *P* values calculated by Mann-Whitney U test for continuous variables; Chi square test or Fisher's exact test for categorical variables.

Table 5. Results of COX1 and COX2 model with determined factors on 5-year OS of female CRC patients based on the training set (n = 58). These two models were only for reference and comparison. They were invalid for interpretation or making predictions because they violated the proportional hazard assumption. Additionally, a small sample size with limited death events led to extremely high HRs. Thus, HRs were invalid and biased.

Model	Variables	HR (95% CI)	*P*	Variable violated PH assumption	Model violated PH assumption	Model C-index (Se)
COX1	glutathione	0.50 (0.20, 1.21)	0.123	No	Yes	0.949 (0.023)
	glutathione disulfide	1.08 (0.48, 2.44)	0.852	Yes		
	glycerol 3-phosphate	8.92 (1.17, 68.22)	0.035	Yes		
	phosphoenolpyruvate	1.26 (0.60, 2.65)	0.544	Yes		
	succinate	0.02 (0.00, 1.68)	0.084	No		
	UDP-D-Glucose	1.23 (0.67, 2.25)	0.504	No		
	Tumor location: RCC	1.09 (0.14, 8.54)	0.936	No		
	Clinical stage: late	66.29 (3.51, 1251.58)	0.005	No		
	Age	1.27 (1.08, 1.49)	0.004	Yes		
COX2	glutathione	0.58 (0.39, 0.88)	0.009	No	Yes	0.952 (0.020)
	glycerol 3-phosphate	9.43 (1.52, 58.55)	0.020	Yes		
	succinate	0.08 (0.01, 0.71)	0.020	No		
	Tumor location: RCC	1.60 (0.27, 9.55)	0.610	No		
	Clinical stage: late	48.48 (4.87, 482.10)	<0.001	No		
	Age	1.25 (1.08, 1.44)	0.003	No		

Table 6. Variables with relative importance larger than 20% using RSF based on the training set (58 female patients). VIMP: variable importance.

Variables	VIMP	Relative importance
Citrulline	0.0228	100.00%
Glutathione	0.0185	81.14%
Asparagine	0.0117	51.45%
LysoPC(16:0)	0.0104	45.60%
Clinical Stage	0.0096	41.88%
Creatinine	0.0079	34.62%
Glucosamine 6-phosphate	0.0078	33.97%
Age	0.0073	31.88%
Glycerol 3-phosphate	0.0057	24.84%
Taurine	0.0049	21.38%
Hypoxanthine	0.0047	20.69%

Table 7. Comparing predictive performance for models in both training set and testing set. COX1 model adopted statistically significant metabolites in individual analysis with clinical stage, tumor location, and age. COX2 model included remaining variables generated from COX1 after forward stepwise selection procedure with minimum AIC. Cox-LASSO was a Cox regression model using all the variables with lasso penalty for feature selection. RSF1 was constructed by the stepwise RSF backward algorithm with a minimum prediction error rate. RSF2 used variables with VIMP > 0.005.

Models	C-index Estimates (95% CI)	
	Training set (n = 58)	Testing set (n = 37)
COX1	0.9494 (0.9043-0.9945)	0.6000 (0.3656-0.8344)
COX2	0.9448 (0.8938-0.9958)	0.6370 (0.4257-0.8483)
Cox-LASSO	0.7056 (0.6903-0.7210)	0.6440 (0.6321-0.6558)
RSF1	0.8117 (0.8104-0.8131)	0.7765 (0.7756-0.7773)
RSF2	0.8469 (0.8462-0.8476)	0.6589 (0.6578-0.6600)
LR	0.7732 (0.6090-0.9375)	0.6125 (0.3686-0.8564)

List of Figures

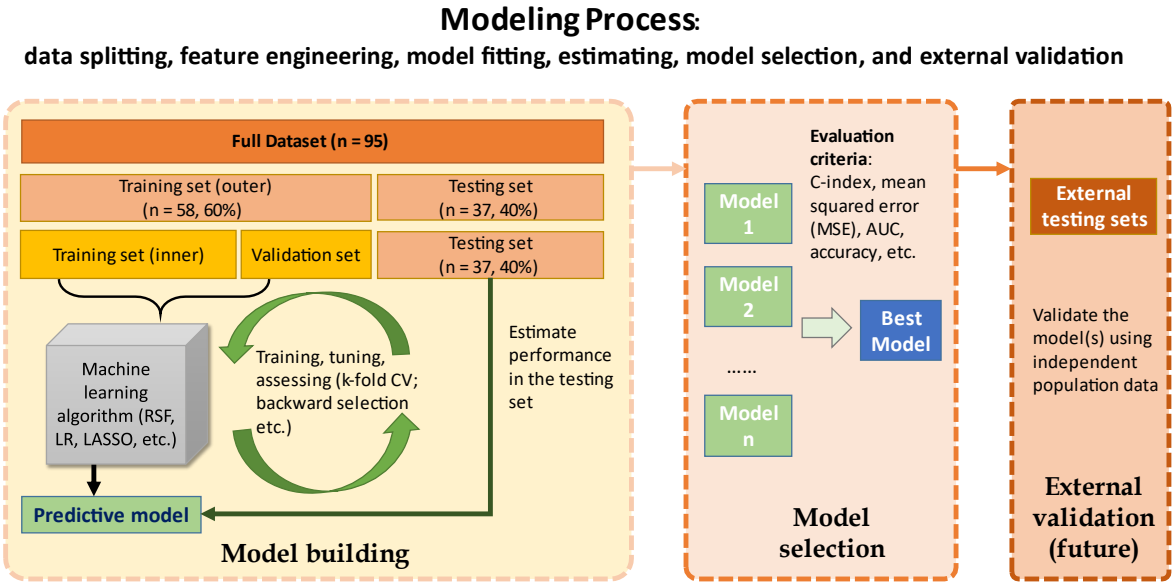
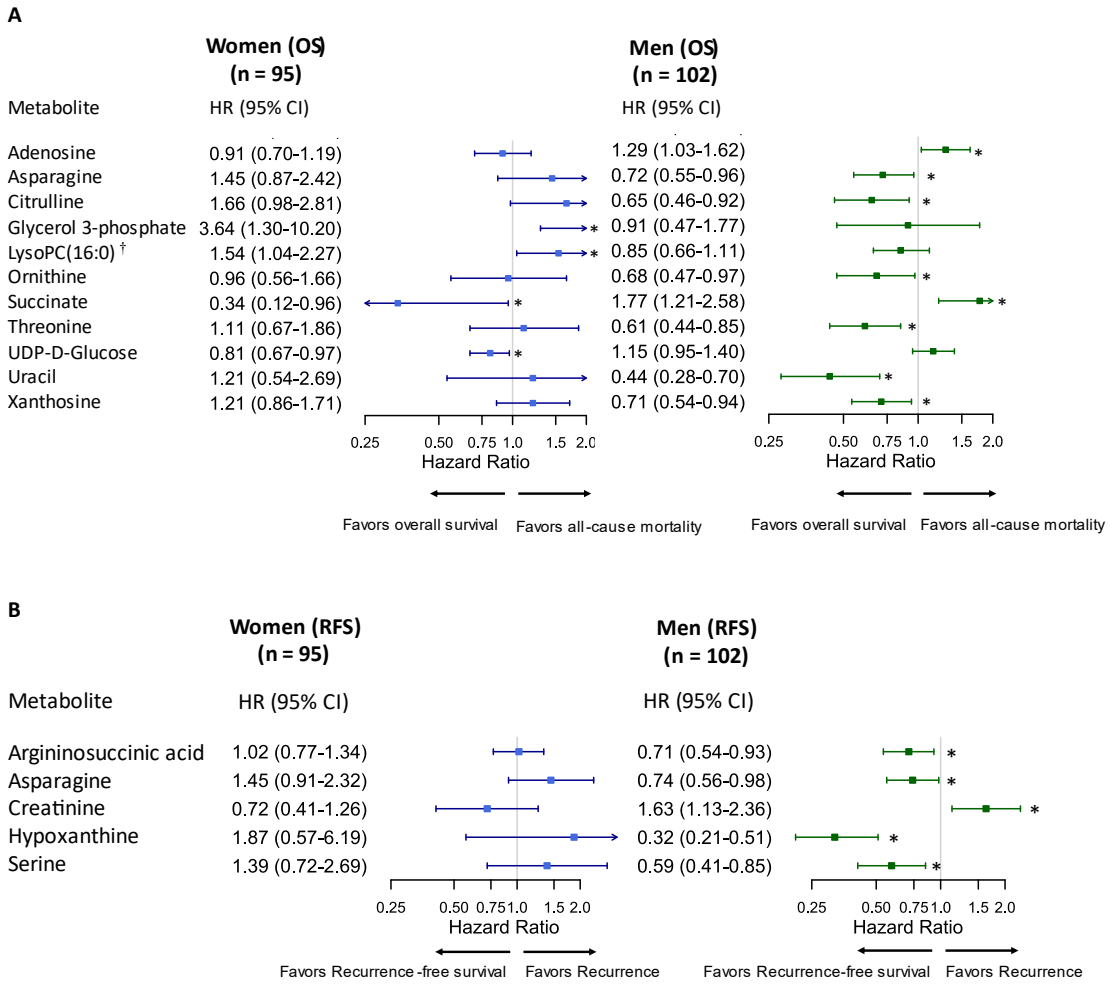


Figure 1. **Flowchart for the modeling process.** RSF: random survival forest. LR: logistic regression. LASSO: least absolute shrinkage and selection operator. K-fold CV: k-fold cross-validation. C-index: concordance index. External validation: an important procedure to ensure sufficient robustness and generalizability using independent external data sets after the models are tested to be valid from internal validation (the current data set we have).



† LysoPC: Lysophosphatidylcholine.

Figure 2. Sex differences in metabolites associated with CRC prognosis. Hazard ratios (HRs) for CRC by sex for individual metabolites (per 1 standard deviation, adjusted for anatomic location, clinical stage, and age (continuous)) (A) 5-year Overall survival (OS) and (B) 5-year Recurrence-free survival (RFS). A metabolite with HR <1 was associated with a protective effect on prognosis; metabolite with HR > 1 was associated with an adverse effect on prognosis. Metabolites with confidence intervals (CIs) marked with asterisks were significantly associated with the presented prognosis (Raw *P* values < 0.05). All the metabolites abundance (continuous) was $\log_2$ transformed. The x-axes are log-scaled. Sex interaction *P* values < 0.05.

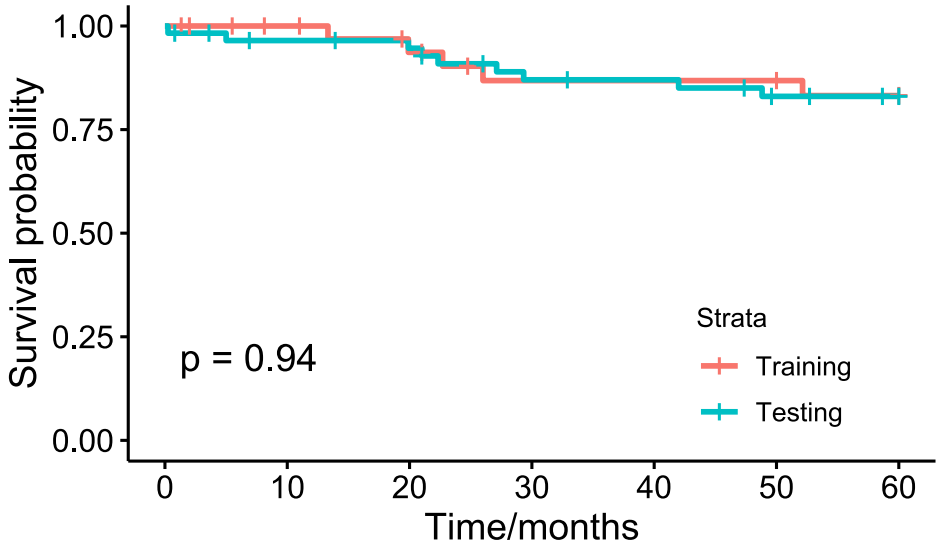


Figure 3. Kaplan-Meier curve of training and testing sets. There was no statistically significant difference between the survival of training and testing sets in log-rank test (P = 0.94).

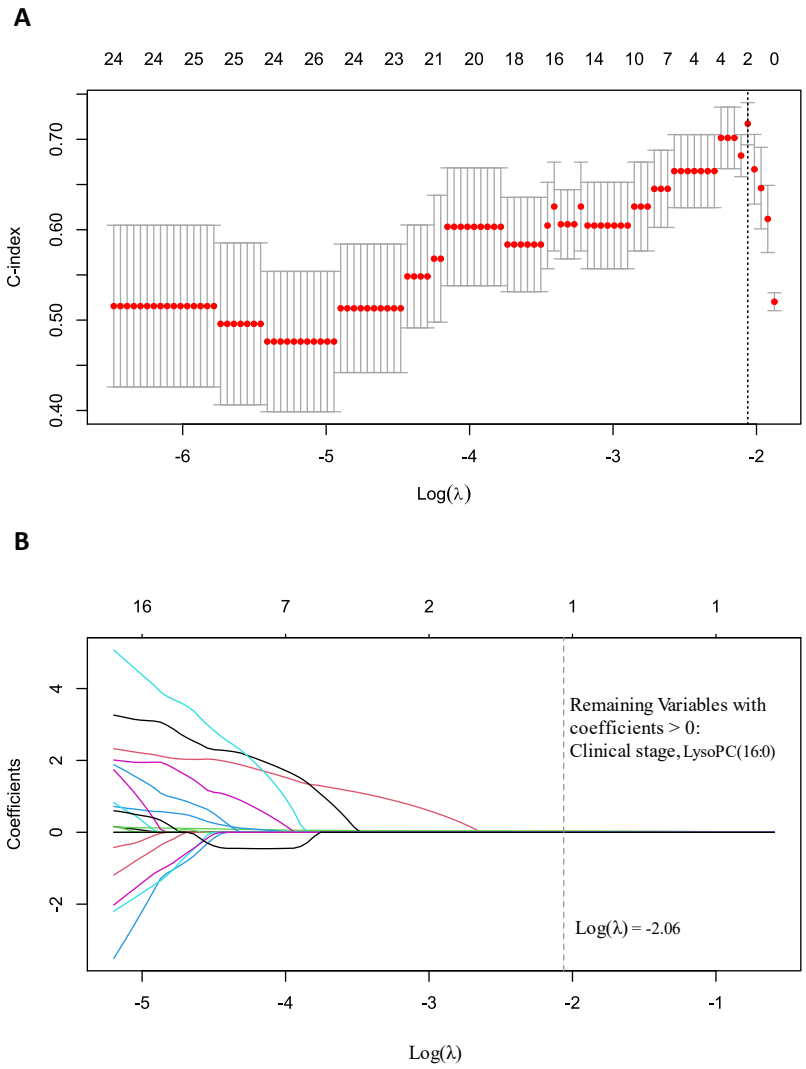


Figure 4. **Cox-LASSO modeling based on the training set (58 female patients).** (A) 10-fold cross-validation curve using the training set as a function of the λ with upper and lower standard deviation bar. The optimum λ corresponded to the highest C-index. (B) a coefficient profile plot of the coefficient paths for a fitted Cox-LASSO model using the training set. At Log(λ)= -2.06, only coefficients of two variables were not penalized to 0: Clinical stage (coefficient = 0.3486) and lysoPC (16:0) (coefficient = 0.2178).

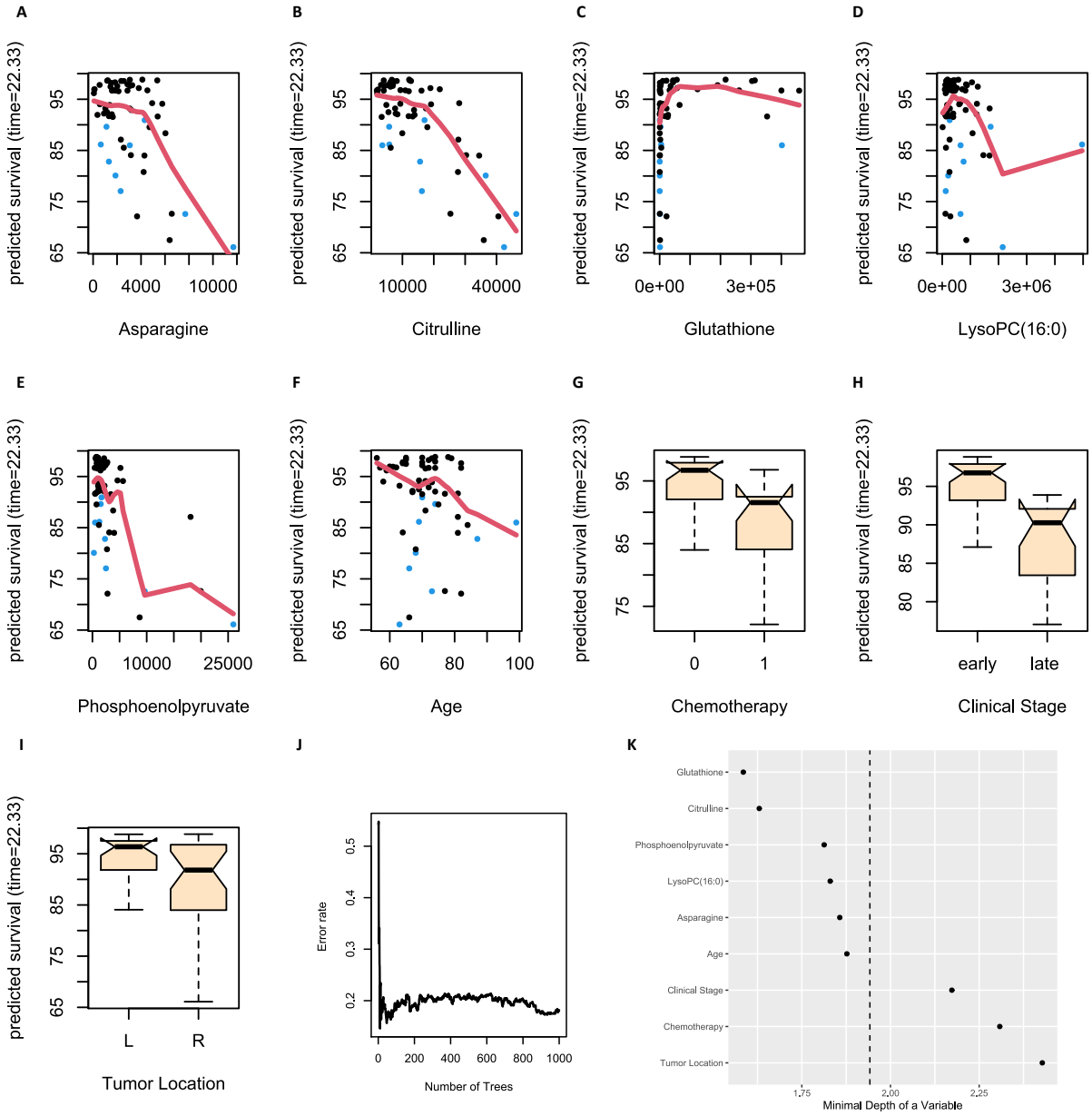


Figure 5. **Partial 5-year predicted survival for nine most predictive variables on survival in colorectal cancer data based on the training set (58 female patients).** Values on the vertical axis represent predicted survival probability for a given predictor after adjusting for all other predictors (A-I). (J) Error rates of RSF for log-rank splitting rule. (K) Identified metabolites that are most predictive for 5-year OS among females by the minimal depth measurement. Metabolites were identified using the random survival forest backward algorithm. Metabolites with lower minimal depth values are more predictive regarding 5-year OS. Abbreviations LysoPC: Lysophosphatidylcholine. Dark dots in A-F represented survivors, and blue dots represented dead patients.

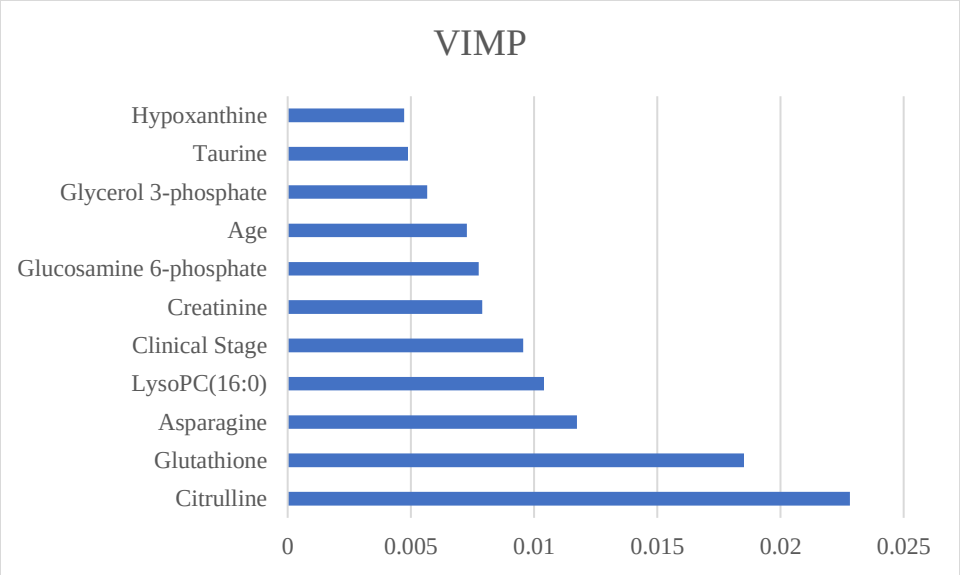


Figure 6. Variable importance (VIMP) for variables with relative importance larger than 20% based on the training set (58 female patients)

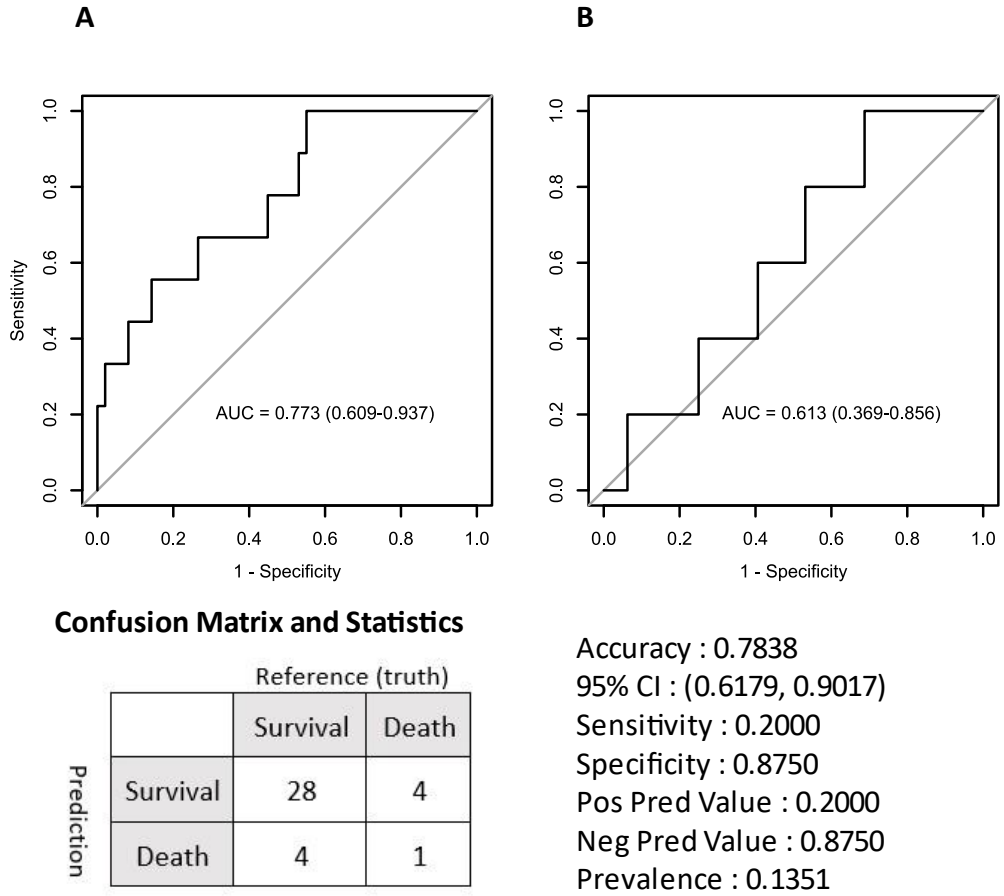


Figure 7. **Prediction performance for Logistic Regression (LR).** ROC curve for training set (n = 58) (A) and testing set (n = 37) (B). ROC curve: receiver operating characteristic curve. AUC: Area under the ROC Curve. Pos Pred Value: Positive prediction value. Neg Pred Value: Negative prediction value.

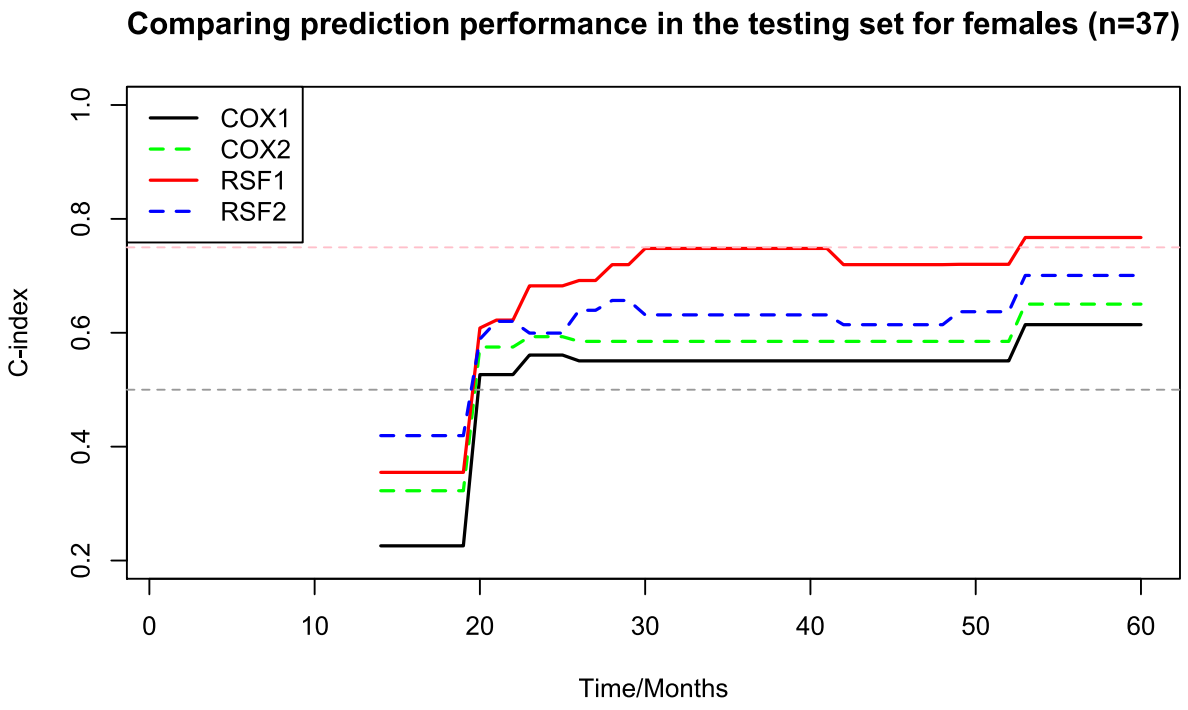


Figure 8. Prediction performance over time for the four models in the testing set (n = 37). RSF1 model had the best predictive performance since month 20, followed by RSF2, COX2 and COX1. The dotted pink line indicates a C-index = 0.75 as an acceptance threshold for a valid predictive model. The dotted grey line indicates a C-index = 0.5. Models with C-index < 0.5 were considered no better than predicting an outcome than random chance. RSF1 reached above the acceptance threshold after month 53.

Introduction

According to CDC reports, colorectal cancer (CRC) is the third major cause of cancer-related deaths in the United States in 2020[1]. CRC survival is not only related to the stage at diagnosis but is also strongly affected by the implementation of population-based screening that reduces colorectal cancer incidence and mortality[2]. Studies have shown that conventional risk factors included aging, family history of cancer, obesity, diets, alcohol consumption, smoking, low physical activity, and socioeconomic status[3]. Clinical and pathological variables such as inflammatory response, body mass index (BMI), tumor location and size, metastasis, lymph node metastasis, and pathological stage of tumors also greatly influence CRC survival and have been incorporated into prognosis prediction[4]. An emerging amount of new biomarker-based approaches have been applied in colorectal cancer screening programs due to less invasiveness, lower costs, and potentially higher detection accuracy, such as DNA methylation biomarker test using blood samples[5] and tumor endothelial marker test[6]. In addition, stratified screening programs by risk factors including sex allow screening and preventive strategies to be targeted at those most likely to benefit while reducing the number of patients undergoing harmful or invasive tests, which unleash the potential to improve screening efficiency[78].

Sex-related differences in CRC prevalence, prognosis, clinical stage, and metabolism have become increasingly popular in cancer research[91011]. Males have poorer survival for CRC[12]. However, females have a higher prevalence of right-sided colon cancer (RCC)[12], which was associated with and poor overall survival (OS)[1314]. Johnson Lab investigated untargeted metabolomics on tumor tissue, looked at the biological mechanisms, and found a positive association between aberrant metabolism in asparagine synthetase (*ASNS*) expression and poor survival[15], which laid the foundation of this study. These findings suggest that sex plays a vital

role in CRC prognosis together with the influence of anatomic tumor location, clinical stage, and metabolism. Thus, it is promising to gain a complete view of how sex interacts with CRC prognosis by addressing the clinical problem from the perspectives of metabolomics that helps to reveal the biological background. Therefore, this study hypothesized that the untargeted metabolic profile of primary tumor tissue metabolites could reveal sex differences in the associations with CRC prognosis, and it could be used to predict CRC prognosis by sex.

A couple of studies have built robust prediction models for CRC prognosis in recent years using clinical factors, biomarker data, and histopathological image data. Roshanaei et al. validated a random survival forest (RSF) model to identify important risk factors (metastasis to other organs, WBC count, disease stage, and the number of lymphomas) on mortality in CRC patients based on their demographic and clinical-related variables[16]. Xu et al. used logistic regression (LR) to predict the recurrence of stage IV CRC after tumor resection by considering time-to-event outcome as a binary outcome (whether recurrence occurred)[17]. Bychkov et al. developed an image-based deep learning approach to predict colorectal cancer outcomes based on images of tumor tissue samples that outperform an experienced human observer in extracting more prognostic information[18]. Kather et al. confirmed that convolutional neural networks (CNN) were able to assess the human tumor microenvironment and predict prognosis directly using histopathological images[19]. Biomarker-based prediction models such as the Circulating free DNA (cfDNA) -based prognostic prediction model based on LASSO-Cox methods achieved an excellent discriminating ability[20].

There are multiple statistics and machine learning approaches available to build models for prognosis prediction in clinical practice. In common, the Cox proportional hazards (CPH) models are used to evaluate the relationships between cancer prognosis and risk factors.

However, CPH is sometimes not suitable for analyzing data with high dimension, complex interactions between variables because it assumes that the outcome is a linear combination of covariates[21]. The proportional hazard assumption is often violated in some survival data[22] , which could produce biased hazard ratios. To avoid the defects of CPH, other non-parametric models are more appropriate in this scenario. For metabolite data specifically, the exploratory analysis of high dimensional metabolomic data containing hundreds of highly correlated variables using regression approaches has unique statistical challenges related to multiple testing and multicollinearity, which had been a major difficulty in this study. Multiple studies have demonstrated that random survival forest (RSF) was a promising approach for identifying disease-associated variables in complex time to event data with a large number of highly correlated metabolites by utilizing a set of decision trees for prediction and ranking variables by their importance[232425]. With RSF backward elimination procedure, Dietrich et al. successfully extracted a series of informative metabolites for predicting type 2 diabetes mellitus (T2D) 5-year disease-free survival, some of which showed nonlinear relationships with prognosis, indicating the necessity of using RSF instead of CPH[23]. Likewise, our study also hypothesized that we could find predictive metabolites for CRC prognosis using advanced statistical methods.

This study is a follow-up to the recent study at Johnson Lab, where we saw that high expression of genes encoding metabolic enzymes were associated with poorer survival in females with RCC[15]. We first looked at tumor tissue metabolites with sex differences in their associations with CRC prognosis (5-year overall survival and 5-year recurrence-free survival) considering anatomic tumor location, clinical stage, and age at diagnosis. Moreover, we examined the possibility of making predictions for CRC prognosis using tumor tissue metabolome considering sex difference, then identified predictive metabolites based on the RSF model for 5-year OS

among female patients. Finally, we built several predictive models, compared their predictive performances, and obtained the optimal one.

Methods and materials

Sample Collection and Metabolites Measurements

Metabolites were extracted and analyzed by hydrophilic interaction chromatography mass spectrometry (HILIC-MS) and reverse phase liquid chromatography mass spectrometry (RPLC-MS)-based metabolomics as previously described in an article by Cai et al.[15]. Only tumor tissue samples from RCCs and LCCs within stage I-III (n=197) were selected in this study. Finally, abundances of 91 metabolites were obtained.

Statistical Analysis

We included age, sex, anatomic tumor location, and clinical stage as covariates. Multivariable Cox proportional hazard regression models were constructed to evaluate the associations between prognosis with both individual metabolite abundance (1 SD differences on a log-scale) adjusted for covariates for all patients and for both sexes. Two prognostic outcomes were considered: 5-year overall survival (OS), 5-year recurrence-free survival (RFS). Due to the absence of death events among females at clinical stage I, we recoded the clinical stages I and II as “early stage”, and III as “late stage”. Patients with any types of chemotherapy prior to the follow-up were coded as having chemotherapy history. Survival analyses were conducted using package “survival” in R (version 4.0.4).

Random survival forest (RSF)

An RSF is computed by a cluster of binary decision trees that have been frequently used to select the most important variables linked with time to event[26]. Minimal depth measurement is implemented to assess how informative a variable is regarding the time until event[23]. Harrell's concordance index (C-index) is equal to 1- prediction error rate, which is commonly applied to evaluate the predictability of a model. Random survival forest models were trained using the "RandomForestSRC" R package. The RSF parameter number of trees and number of node splits were fixed at 1000 and 10 initially. We applied a random survival forest backward selection algorithm for variable selection to detect the most predictive and informative metabolites while forcing covariates into our models[23], which finally automatically chose the set of metabolites producing the lowest prediction error rate. We used raw abundance in RSF modeling.

LASSO-based CPH

Regression with LASSO (least absolute shrinkage and selection operator) penalty is a commonly used method for variable selection in a high-dimensional data analysis that produces results depend on the shrinkage parameter λ[27]. The R package "glmnet" was applied to Cox-LASSO modeling based on 10-fold cross-validation.

Logistic Regression (LR)

Logistic regression (LR) can predict the probability of occurrence of an event by fitting data to a sigmoidal S-shaped logistic curve[28]. Unlike survival models, LR dropped the time information and coded the event within a certain length of time as a binary variable (e.g., survival during the 5-year of follow-up = 0, death within 5-year follow-up = 1).

Modeling process

We ran RSF based on the backward selection algorithm for all patients and by sex for 5-year OS and RFS and then calculated minimum prediction error rates (MPER) for these six models that were forced to include clinical stage, age, anatomic tumor location, chemotherapy as covariates. MPER lower than 25% was considered as a standard indicating potential good predictive ability[16], and we only further investigate models with MPER < 25% in our study.

The whole modeling process could be summarized in Figure 1. To obtain a model with high generalizability, it is essential to split the data set into a training set and a testing set. The training procedure was conducted using an "inner" training set and validation set if we adopted the k-fold cross-validation technique for parameter tuning based on the machine learning algorithm of our choice. After multiple training cycles, we achieved a model with high fitting performance for the "outer" training set with 60% of the samples. If the performance was poor, we considered it as underfitting, and we would not proceed with further testing on the testing set. In this study, we aimed to build models with C-index at least over 0.75 (equal to prediction error rate < 0.25) in the training set (outer) that represented a promising potential of good predictive ability and then to test them in the testing set. If the C-index for the fitted model on the testing set is high (> 0.75), the model is robust and generalizable and would be considered for external validation using data from other independent cohorts. Otherwise, if the predictive accuracy is not high enough, more valuable information should be collected and analyzed in future modeling for improvement, and combination with other screening, testing approaches would be necessary as a supplementation in clinical practice.

Results

Clinical characteristics

Baseline characteristics of 197 CRC cases (102 males and 95 females), including 5-year overall survival (OS) and 5-year recurrence-free survival (RFS), are shown in Table 1. In the cohort, 37 total deaths and 30 recurrences were documented. The median follow-up time since the date of surgery for primary tumor was 74.8 months (range, 0.1-169.2 months). Older age at which the surgery was performed, having chemotherapy history, and advanced clinical stage were inversely related to OS survival rate. For each subgroup by anatomic tumor location and clinical stage, the demographic characteristics are displayed in Supplementary Table 1. Prognosis among different subgroups is shown in Supplementary Table 2. Among these variables, chemotherapy history was significantly associated with clinical stage for all patients (Wilcoxon rank-sum test P value < 0.001), and the treatment effects of chemotherapy on prognosis counteracted the harmful effects of being late stages, which would make models hard to interpret. Thus, we believed that clinical stage provided enough information, and we did not include chemotherapy history as a covariate in the Cox Proportion Hazard (CPH). But we used both variables in other models using machine learning algorithms that could carry feature selection automatically and produce interpretable results.

Cox proportional hazard regression (CPH) analysis

In Table 2, sex was significantly associated with 5-year OS adjusted for anatomic tumor location, clinical stage, age. We then examined the sex differences in the associations between OS and metabolome and whether it was necessary to build different models by sex for prognosis

prediction. Sex seemed to be independently associated with RFS but still conducted the same analysis for RFS to see if any metabolites had sex heterogeneity in the associations with RFS.

We first analyzed the relationships between the abundance of 91 metabolites and OS. Multivariate Cox proportional hazard (CPH) models estimated associations between 91 tumor tissue metabolites and CRC prognosis individually by sex with 1 SD differences (log-metabolite scale), adjusted for anatomic location, clinical stage, and age: 18 metabolites were significantly associated with OS (Supplementary Table 3) for either female or male patients (Supplementary Table 4). Twenty-five metabolites had statistically significant correlations with RFS for either females or males (Supplementary Table 4). For both supplementary tables 3 and 4, the *P* values were raw values before FDR adjustment. Only carnitine and hypoxanthine remained significantly associated with RFS for males after FDR adjustment. Fig. 2 summarizes metabolites whose associations with CRC prognosis differed by sex (interaction *P* values < 0.05).

Adenosine, asparagine, citrulline, glycerol 3-phosphate, LysoPC (16:0) (lysophosphatidylcholine (16:0)), ornithine, succinate, threonine, UDP-D-Glucose, uracil, and xanthosine were found to have significant sex differences in their associations with CRC OS (Supplementary Table 3, Fig. 2A). Among these 11 metabolites, succinate was associated with better OS for females (HR_{OS}=0.34 per SD, 95% CI: 0.12-0.96, *P* = 0.042), while it was associated with poorer overall survival for males (HR_{OS}=1.77 per SD, 95% CI: 1.21-2.58, *P* = 0.003) (Fig. 2A). Argininosuccinic acid, asparagine, creatinine, hypoxanthine, and serine were found to have significant but opposite associations with RFS between female and male patients (Supplementary Table 4, Fig. 2B). These metabolites were all significantly associated with RFS in males but were not associated with RFS in females. Interestingly, asparagine was observed to have sex differences for both OS and RFS (Fig. 2). Asparagine was significantly associated with better

CRC prognosis in male patients (both OS and RFS): HR_{OS}=0.72 per SD, 95% CI: 0.55-0.96, P = 0.025; HR_{RFS}=0.74 per SD, 95% CI: 0.56-0.98, P = 0.039, but there were no significant trends in female patients (interaction P_{OS} = 0.029, interaction P_{RFS} = 0.009) (Supplementary Table 3, 4). None of the results in Fig. 2 violated the proportional hazards assumption.

Multivariate CPH analysis that includes all metabolites with clinical variables was inappropriate because of strong collinearity and divergent results, given relatively small sample size and high dimension with around 100 variables. Thus, we hoped to reduce dimension by Principal Component Analysis (PCA). We tried to implement CPH analyses combined with PCA, but there was no statistically significant result for OS among female CRC patients (Supplementary Fig. 1, Supplementary Table 5, Supplementary Table 6). Thus, we sought other methods to identify predictive metabolites.

Predictive modeling

Primary exploration of the possibility to predict CRC prognosis using tumor tissue untargeted metabolic profile

The identified sex interactions indicated potential sex heterogeneity. Thus, we conducted predictive modeling taking sex differences into account. We found that CPH models that included all the metabolites to select features could not converge for all patients or both sexes, even with stepwise selection methods. So, we turned to the Random Survival Forest algorithm (RSF) to handle high dimensional data and collinearity problems without having to consider the proportional hazard assumption. We ran RSF based on the backward selection algorithm for all patients and by sex for 5-year OS and RFS and then calculated prediction error rates (PER) of the model with selected variables together with clinical stage, age, anatomic tumor location, chemotherapy as covariates. We further investigated models with minimum PER lower than

25%, which was considered as a standard indicating potential good predictive ability in our study. The selected metabolites were different by sex.

For example, the minimum prediction error rate for all patients was 0.3085 (Table 3). Thus, it was not good to predict OS for all patients using our metabolomic profile based on RSF. It seemed that predictive modeling using tumor tissue metabolome based on RSF for OS in females and RFS for males was plausible with a minimum PER of 0.2100 and 0.2356, respectively (Table 3). To justify the model, it is necessary to follow a standard machine learning modeling process that split the data into a training set and testing set, and picked a trained model based on the training set and tested it on the testing set (Fig. 1). It is worth mentioning that the training set results might not be the same as results from the entire data set for 95 female CRC patients due to sampling randomness. Furthermore, a successful random split would not cause too extreme deviations between the two results. So far, this step was only meant to explore the possibility of making predictions for different prognostic outcomes and populations. We could make inferences using the results, but it would be better to test the findings in an independent data set. Since we only obtained one cohort data, the only way is to split it into two independent parts to for modeling, as Fig.1 described. We used a training set with 60% randomly selected patients to select features and test it in a test set consisting of the remaining 40% patients.

For simplicity, in this thesis, we only further investigated models for OS for female patients with a lower prediction error rate of 0.21 using the RSF algorithm. The entire dataset was split into two independent groups, 60% for training and 40% for testing. There were no statistically significant differences among the features of the two groups (Table 4). Mann-Whitney U test for continuous variables; Chi-square test and Fisher's exact test for categorical variables. The difference in survival outcome was absent between the two sets as well (Fig. 3) so that the death

events were balanced. Then it was safe to use the training set to generate a prediction model and employ the testing set to estimate the model's accuracy.

Cox proportional hazard regression (CPH) models with variable selection using forward stepwise and LASSO

We built the first CPH model (COX1) with statistically significant metabolites in the individual analysis with clinical covariates from the training set: glutathione, glutathione disulfide, glycerol 3-phosphate, phosphoenolpyruvate, succinate, UDP-D-Glucose (Table 5). Then we used forward selection methods to select a group of variables with the least AIC (Akaike information criterion) for CPH model (COX2) that included clinical stage, age, succinate, glycerol 3-phosphate, and glutathione (Table 5). Recurrence status was correlated with OS (Fisher's Exact Test: $P < 0.001$), but it should not be included in any of the models since this was observed simultaneously with OS in practice. For COX1, the C-index was 0.9494 (0.9043-0.9945) for training set and 0.6000 (0.3656-0.8344) for testing set as shown in Table 7; for COX2, C-index was 0.9448 (0.8938-0.9958) for training set and 0.6370 (0.4257-0.8483) for the testing set. However, both models violated the proportional hazard assumption, which might be due to nonlinear covariate relationships or lack of independence that made the results less reliable. These two models in Table 5 were only for illustration and reference; they were invalid for making predictions and interpretation because the violations of the proportional hazard assumption produced biased hazard ratios.

The variable selection methods for both COX1 and COX2 were primitively conducted manually, and thus we turned to other methods suitable for addressing dimensionality reduction. Cox-LASSO regression could use L1 penalty for feature selection and dimension reduction. Using 10-fold cross-validation (CV), we tuned the λ parameter and selected the best one to produce a

selected set of variables. We trained the Cox-LASSO in our training set, as shown in Fig. 4, where the C-index peaked at 0.7173 when λ was around 0.1273 (Log λ = -2.06). When λ is very small, the LASSO produced results similar to a CPH model with all the coefficients included. As λ grew, the regularization term had a greater effect on penalizing more variable coefficients to zero, leaving fewer variables in the model. Finally, only variables that were influential enough were included in the model. The training process was summarized in Fig. 4. At Log(λ)= -2.06, only two variables had coefficients > 0: Clinical stage (coefficient = 0.3486) and lysoPC (16:0) (coefficient = 0.2178), both of which were positive indicating their contribution to lower overall survival probabilities. We used this λ to test the prediction performance on the testing set, and we got a C-index of 0.6667. To obtain a mean and 95% CI for the C-index for the training set and testing set, we ran the model 1000 times, and the C-index was 0.7056 (0.6903-0.7210) for the training set and 0.6440 (0.6321-0.6558) for the testing set as shown in Table 7.

Random survival forest (RSF) model

We first used the stepwise RSF backward algorithm by selecting a group of variables that produced the best prediction performance using the training set. The model (RSF1) included asparagine, citrulline, glutathione, lysoPC (16:0), phosphoenolpyruvate together with forcibly included covariates (tumor stage, clinical location, age, and chemotherapy history), and the prediction error rate reaches a minimum of 0.1883 (C-index=0.8117). Then we built an RSF model (RSF2) including variables with relative variable importance (VIMP) > 5% (RSF1) for both females: clinical stage, citrulline, chemotherapy history, age, hypoxanthine, glycerol 3-phosphate, glutathione, asparagine, DHAPorG3P, and spermine (Table 6, Fig. 6). We ran the models 1000 times to obtain their mean and 95% CI. RSF1 had a high C-index in the training set of 0.8117 (95% CI: 0.8104-0.8131) and a lower C-index for the testing set 0.7765 (95% CI:

0.7756-0.7773). RSF2 had a high C-index in the training set of 0.8469 (95% CI: 0.8462-0.8476), but its C-index for the testing set is much lower (0.6589, 95% CI: 0.6578-0.6600).

The associations between the nine most predictive variables by RSF1 and OS are demonstrated in Fig. 5. The estimated partial survival for a covariate indicates estimated survival for different levels of the covariate after accounting for the average effects of the other selected metabolites and the covariates. It can be seen from Figure 5, in continuous risk factor for example, as the asparagine abundance increased up to about 4000, the 5-year predicted overall survival probability decreases slowly from 95% to 90%, and then it decreased at a sharper rate until reach 65% (Fig. 5A), and a similar trend was found in citrulline (Fig. 5B). Glutathione, lysoPC (16:0) and phosphoenolpyruvate had nonlinear relationships between predicted survival probability their abundances with several turning points in their plots. For categorical variables such as anatomic tumor location, right-sided cancer (RCC) demonstrated a lower 5-year predicted survival estimated at around 5% compared with LCC on average (Fig. 5I), which agreed with many previous findings[2930]. Female patients with chemotherapy history or at the late clinical stage had about a 5% lower probability of survival. As Fig. 5J shows, the prediction error rates decreased drastically and became stable and stayed around 0.18 as the number of trees grew to 1000. It is worth mentioning that calculation methods for these partial survival plots were not the same as Kaplan–Meier curve or CPH models, and RSF models do not have to observe the proportional hazard assumption.

Logistic regression model (LR)

Logistic regression (LR) was not the same as other methods that treated the data as time-to-event data. Instead, it neglected the survival time and coded the death event as a binary variable

(survival = 0, death = 1). The methods simplified the question into predicting "whether the patient would die or not within the 5-year interval" without considering when the death would occur. The predictive performance in the training set was better than the testing set, which indicated a typical overfitting problem within the LR method (Fig. 7): LR had good specificity (0.875) but inferior sensitivity (0.200). High specificity demonstrated this method had a strong ability to correctly designate an individual who would not die in 5 years as a survivor, which would help avoid unnecessary financial costs or mental burden for CRC patients. Low sensitivity corresponded to more false-negative results, and thus more events of death within 5-years would be not be anticipated, resulting in losing opportunities for early preventive intervention for CRC female patients. The low sensitivity made its AUC (area under the curve) for the testing set relatively low with a lower bound below 0.5. As a result, this method should be improved before being used in clinical practice.

Comparing predictive performance for the models

The performances of these five models were evaluated based on C-index (1-prediction error rate). A C-index larger than 0.75 is desired and indicates a good model. C-index < 0.5 indicates a poor performance meaning that the model is no better than predicting an outcome than random chance. 95% CI that includes 0.5 is considered to be not significant. For RSF1, RSF2, and Cox-LASSO, 95% CIs were calculated by running the models in the testing set 1000 times. COX1 and COX2 fit the training data well with less error rate than other methods. However, their predictive abilities in the testing set were not satisfying, with a lower bound of 95% CI below 0.5 (Table 7). Again, COX1 and COX2 were invalid and were just for comparison because of biased hazard ratios due to violations of the proportional hazard assumption. RSF1 model outperformed in predicting female 5-year OS with two stable and relatively high C-indexes of 0.8117 (95% CI:

0.8104-0.8131) and 0.7765 (95% CI: 0.7756-0.7773) for the training set and the testing set, respectively; both were above the acceptance threshold of 0.75. Results show an overfitting problem with RSF2 as its performance in the training set (C-index = 0.8469, 95% CI: 0.8462-0.8476) was good, while its prediction error in the testing set was huge (nearly 0.34). Cox-LASSO had problems of underfitting with the lowest C-index for the training set among the six models (C-index = 0.7056, 95% CI: 0.6903-0.7210), though the performance was more stable than LR. For future clinical practice purposes, RSF1 has the potential to be accepted if updated using other established clinical and biological variables combined with other screening measurements. Other models suffered from either underfitting or overfitting with a lack of robustness in the testing set.

The prediction performances over time of COX1, COX2, RSF1 and RSF2 were examined as shown in Fig. 8. The predictive abilities of the four models in the testing set were examined over time (the testing set did not have a death event before month 20). All four models showed better predictive abilities at a later time. RSF1 had a stably good prediction performance since month 30 in general. C-index of RSF1 during month 30 to month 40 hit 0.75, then dropped a little bit before month 53, and soon reached above 0.75, indicating a promising possibility of predicting OS at any time during 3-5 years.

Discussion

This study is the first cohort using untargeted metabolomics to investigate possible associations between 91 tumor tissue metabolites and colorectal cancer prognosis. Our analyses on individual metabolites identified 11 metabolites with sex interactions in the associations with colorectal cancer 5-year overall survival (OS) and five metabolites for 5-year recurrence-free survival (RFS), among which asparagine was observed to have sex dimorphisms for both OS and RFS. These findings suggest that different metabolism by sex were associated with different CRC prognostic outcomes, and it was vital to build targeted predictive models by sex from the point of view of precision medicine. By applying an RSF backward selection procedure within the training set for female CRC patients (n = 58), five metabolites were identified to be most predictive for the 5-year OS: glutathione, citrulline, phosphoenolpyruvate, lysoPC (16:0), and asparagine. As demonstrated by the RSF1 model that incorporated these five metabolites might provide new insights to the prediction of colorectal cancer 5-year OS for female patients when used together with known epidemiological risk factors of CRC (clinical stage, chemotherapy history, anatomic tumor location, and age). The comparison of the C-index (1- prediction error rate) of five other different models revealed that especially noise metabolites were removed by the RSF backward selection process, resulting in identifying the most predictive metabolites. In contrast, RSF2 that included relative variable importance > 20% showed a better fitting performance in the training set than RSF1, but it could not achieve a comparable C-index in the testing set, which suggested its poor generalizability due to overfitting. Moreover, the visualization by partial plots revealed nonlinear associations between the abundance of identified metabolites and predicted 5-year overall survival, indicating possible diagnostic cut points for further research.

Our exploratory analyses and modeling indicated that traditional hazards-based models such as CPH were not designed for prediction but to infer variables' impact on a prognostic outcome[21], and were too primitive for high-dimensional metabolomics data. Instead, machine learning algorithms performed better in predicting prognosis when we faced nonlinear interactions that were presented in Fig. 5, which would violate the linear proportional hazards condition (Table 5). For example, Fig.5 I showed that RSF successfully distinguished the difference in 5-year predicted survival between right-sided cancer (RCC) and left-sided cancer (LCC), which might be hidden under the complex interactions between anatomic tumor location and tumor tissue metabolites. As expected, none of the CPH models discovered the location-specific difference, as CPH could not handle intricate inner interactions between metabolites and clinical variables so explicitly. A similar modeling process to model nonlinear gene interactions made comparisons between CPH and other machine learning methods, including RSF, which also proved the applicability of automatically assessing nonlinear effects and complex interactions by RSF[21]. These data-driven machine learning algorithms are unaffected to problems due to their natures that perform robust feature selection against multicollinearity internally. As a result, collinearity between variables did not impair the predictive accuracy and satisfied our goal of disengaging from multicollinearity problems[21]. In our study, RSF has also shown its ability to outperform classic CPH regressions at any time within the 5-year follow-up. As can be seen in Table 7, the two CPH models COX1 and COX2, had very high C-index in the training set, while they failed to handle the testing set. RSF1 using the backward selection algorithm showed the best performance, with the C-index of the training and testing sets reaching 0.812 and 0.777, respectively. The performance of RSF1 kept at a steady high C-index over time since month 20.

Interestingly, the variables with coefficients larger than 0 in the Cox-LASSO model were lysoPC (16:0) and clinical stage, which were also selected as predictors in RSF1 and RSF2, while COX1 and COX2 did not include the lysoPC (16:0) metabolite. LysoPC (16:0) had a significant sex difference in overall survival in the individual metabolite analysis and was a risk factor for female OS (HR_{OS}=1.54 per SD, 95% CI: 1.04 - 2.27, P = 0.031). Cai et al. found that the lysophospholipids lysophosphatidylcholine were upregulated in women with RCC (stage I) but not in men, suggesting that the higher levels of lysophospholipids in women with RCC would promote fatty acid supply[15] that was essential for cancer cell growth[31]. These findings justified our findings identifying lysoPC (16:0) as an important predictor for CRC prognosis for females.

We also found that asparagine was an essential metabolite for female prognosis. Asparagine had sex interactions with both OS and RFS and was tested to be an important predictor in both RSF1 and RSF2 models. Asparagine (Asn) abundance was associated with lower probabilities of overall survival, which agreed with the previous finding of female survival and asparagine synthetase (*ASNS*) expression[15]. Johnson Lab found that asparagine increased threonine uptake in females RCCS that were nutrient deplete and could lead to aggressive phenotypes in those patients[15]. In Fig.2, asparagine and threonine were all not significantly associated with OS for female patients but were both associated with better OS for males. In cancer research based on *in vitro* experiments, *ASNS* catalyzed asparagine was crucial for cancer cell growth by promoting cancer cell amino acid homeostasis, anabolic metabolism, proliferation[32], and Asn availability *in vitro* strongly interplayed the metastatic progression of breast cancer[33]. For CRC specifically, *SOX12* expression promoted colorectal cancer cell proliferation and metastasis and facilitated *ASNS* expression[34]. Another frequently found mutation in the *KRAS* gene in colorectal cancer[35] was observed with a marked decrease in aspartate level and increased asparagine level by an

upregulated *ASNS* expression, which indicated that *ASNS* might be a novel therapeutic target for *KRAS*-mutant CRC [36]. Moreover, *SLC25A22* served as an essential metabolic regulator for CRC progression by promoting the synthesis of aspartate-derived amino acids (asparagine) in *KRAS*-mutant CRC cells[37]. Knott et al. reported that increased dietary asparagine in animals promotes metastatic progression in breast cancer, and dietary asparagine restriction inhibits metastasis without affecting the primary tumor growth[38]. The study drew great attention from academics and the media worldwide in 2018, ranked the 97th percentile (ranked 24th) of the 896 tracked articles of a similar age in Nature on Altmetric[39]. According to BBC News, researchers from Cambridge University claimed that patients with specific cancers might have developed an addiction to specific components of diets, and it may be necessary to modify a patient's diet or change the way tumor cells get access to those nutrients with potential risks using drugs[40]. Consequently, both internally produced asparagine and external exposure to asparagine may influence CRC tumor progression. These studies provided strong evidence to support our identification of asparagine as a predictive risk factor of OS for female CRC patients.

Besides, application of RSF with backward selection for all 95 female patients revealed that creatinine was a predictive factor for 5-year OS. Creatinine was found to be a valid variable for predicting CRC cases[41] and was also reported with correlations to colon cancer based on other studies of urine[42] and serum[43] samples from colon cancer patients. Also, creatinine is a measure of cachexia, a syndrome characterized by unintentional weight loss[44]. The female patients we studied were all over 55 years old (with an average age of 71), which might allow us to identify creatinine as a predictive metabolite. However, whether creatinine has predictive ability among other age groups requires further investigations.

Some of the metabolites we observed might have multiple uses and are not just used by the cancer cells, such as polyamine, which could be used by both colon cancer cells and bacterial cells to build biofilms in colon adenomas/carcinomas[45]. Among the identified metabolites, asparagine utilization could be regulated by L-glutamine via gut microbiota[46]. Hence if the environmental milieu of the colon or rectum differs between males and females, it could determine how metabolites are used and thus affect cancer progression.

Admittedly, because of the heterogeneity of different metabolomics data, there is no panacea model for predicting CRC based on any types of metabolomics data. The flowchart in Fig. 1 illustrates a flexible, dynamic path for disease-related metabolomics research discovery. An increasing number of biomedical studies utilizing Automated Machine Learning (AutoML) methods have been applied to diseases such as cancer[47], Alzheimer's disease[48], and cardiovascular diseases[49], which leveraged advances in hyperparameter search and model selection based on metabolomics. Those studies used greater numbers of machine learning algorithms and selected the optimal ones that suit their data best. Therefore, chances are that the best model we built might not be the optimal one for our data, though random forest-based models tend to be more stable than the simple decision tree model. Nevertheless, this study could offer some insights into using metabolome to predicting CRC prognosis accounting for sex differences for future studies.

The strength of this study is the application of untargeted metabolomics for CRC tumor tissue in a well-described population-based retrospective cohort with strictly standardized study protocols and a decades-long follow-up time. Tumor tissue metabolites have advantages over biomarkers extracted from blood and urine samples because tumor tissues directly reflect tumor microenvironment and metabolism, whereas components of other biofluids are liable to external

environmental interactions such as dietary intake. Also, tumor tissue samples are easier to store and more reliable for evaluating and predicting long-term prognosis than plasma samples that tend to degrade in hours or days[50]. In addition, Cai et al. discovered that CRC tumor tissue metabolites also differed by anatomic tumor location, and females with RCC were at higher risks of poor overall survival[15], implying that studying sex differences in prognosis should include tumor location. Hence tumor tissue metabolic profiling considering tumor location was the optimum approach.

Nonetheless, we should also admit the huge gap between real-world medical practice and bioinformatics studies due to low reproducibility, even some of which claimed to provide robust models. Evaluation of 184 studies on new prognostic markers of outcomes in acute pancreatitis showed that only 15% had a sample size > 100 patients, and < 40% reported information about patient recruitment, and none had power calculations[51]. Lack of replication efforts, small sample sizes, insufficient subsequent external validation, unclear evaluation criteria were the major causes of the failures in cancer biomarker discovery and translation along the biomarker pipeline[525354]. There is still a long way to go before our findings are applied to actual medical practice for colorectal cancer, such as providing risk scores by measuring tumor tissue metabolic profile. Despite these possible defects and obstacles, this study could provide hints about predictive prognostic biomarkers for colorectal cancer from the perspectives of sex difference and tumor tissue metabolome.

Conclusion

We identified five predictive metabolites for female CRC patients that could be used to predict 5-year overall survival using the random survival forest (RSF) backward selection algorithm. We concluded that the RSF prediction method based on the backward selection algorithm was promising in predicting 5-year overall survival (OS) for female CRC patients. The results could also justify the urgent need for personalized CRC screening programs by sex or other factors which benefit a targeted group of the population at higher risks with optimal resource allocation.

Due to a large number of correlated variables that brought problems of multiple comparison, insufficient statistical power, and higher risks of multicollinearity, false-positive detection with significant *P* values by chance are sometimes unavoidable in exploratory data analysis of complex metabolomic data based on traditional statistical regression approaches. Moreover, most metabolites identified with sex differences in CRC prognosis had insignificant *P* values after FDR adjustment., which need future replication studies to confirm their associations with CRC prognosis and sex differences. Fortunately, we were able to take advantage of bootstrapping and the interruption of intercorrelation structures by random node splitting to reduce overfitting, multicollinearity, and select reliable predictive variables using RSF. Furthermore, nonlinear relationships between the identified metabolites and predicted survival time could be visualized to determine potential clinical thresholds after validated in further population studies.

The predictive performance of this method in the training set was satisfactory (C-index = 0.81), and the prediction accuracy in the testing set was slightly lower but still acceptable (C-index = 0.78). The model is reliable in the statistic aspect but may need further improvement in clinical practice that requires much higher accuracy. Several limitations might lead to these results: 1) a

small sample size with limited death event for data mining could only provide limited information; 2) established covariates for predicting CRC development and poor prognosis were absent in our data set, such as genotype, family history, metastasis, BMI, dietary factors, alcohol consumption, smoking behaviors; 3) incomplete information of the treatment or drug history within the follow-up period. The already good predictive accuracy in the testing set laid the foundation for improving the prediction performance after adding more variables mentioned above using a larger cohort. Moreover, it is also helpful to conduct a multi-omics analysis that is more comprehensive than metabolomics alone. Afterward, several external validation processes using independent cohort data are necessary before being applied to clinical practice. We foresee the enormous potential of using novel biomarkers to predict prognosis by multi-omics approaches based on machine learning, statistical learning methods.

Ultimately, my recommendations for the YSPH MPH program would be to set up new courses focusing on biomarker discovery for cancer epidemiology (both methodology, data analysis, and causal inference).

References

[1] Colorectal Cancer Statistics, <https://www.cdc.gov/cancer/colorectal/statistics/index.htm> (2020).

[2] Centers for Disease Control and Prevention (CDC). Vital signs: Colorectal cancer screening, incidence, and mortality--United States, 2002-2010. MMWR Morb Mortal Wkly Rep. 2011 Jul 8;60(26):884-9. PMID: 21734636.

[3] Lee CH, Cheng SC, Tung HY, Chang SC, Ching CY, Wu SF. The Risk Factors Affecting Survival in Colorectal Cancer in Taiwan. Iran J Public Health. 2018 Apr;47(4):519-530. PMID: 29900136; PMCID: PMC5996318.

[4] Sharma R, Zucknick M, London R, Kacevska M, Liddle C, Clarke SJ. Systemic inflammatory response predicts prognosis in patients with advanced-stage colorectal cancer. Clin Colorectal Cancer. 2008 Sep;7(5):331-7. doi: 10.3816/CCC.2008.n.044. PMID: 18794066.

[5] Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, Song X, Lesche R, Liebenberg V, Ebert M, Molnar B, Grützmann R, Pilarsky C, Sledziewski A. DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem. 2008 Feb;54(2):414-23. doi: 10.1373/clinchem.2007.095992. Epub 2007 Dec 18. PMID: 18089654.

[6] Pietrzyk Ł, Korolczuk A, Matysek M, Arciszewski MB, Torres K. Clinical Value of Detecting Tumor Endothelial Marker 8 (ANTXR1) as a Biomarker in the Diagnosis and Prognosis of Colorectal Cancer. Cancer Manag Res. 2021 Apr 9;13:3113-3122. doi: 10.2147/CMAR.S298165. PMID: 33859497; PMCID: PMC8043785.

[7] Lin OS, Kozarek RA, Schembre DB, Ayub K, Gluck M, Cantone N, Soon MS, Dominitz JA. Risk stratification for colon neoplasia: screening strategies using colonoscopy and computerized tomographic colonography. Gastroenterology. 2006 Oct;131(4):1011-9. doi: 10.1053/j.gastro.2006.08.015. PMID: 17030171.

[8] Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res (Phila). 2016 Jan;9(1):13-26. doi: 10.1158/1940-6207.CAPR-15-0274. Epub 2015 Oct 13. PMID: 26464100; PMCID: PMC7610622.

[9] Yang Y, Wang G, He J, Ren S, Wu F, Zhang J, Wang F. Gender differences in colorectal cancer survival: A meta-analysis. Int J Cancer. 2017 Nov 15;141(10):1942-1949. doi: 10.1002/ijc.30827. Epub 2017 Jun 24. PMID: 28599355.

[10] Majek O, Gondos A, Jansen L, Emrich K, Holleczek B, Katalinic A, Nennecke A, Eberle A, Brenner H; GEKID Cancer Survival Working Group. Sex differences in colorectal cancer survival: population-based analysis of 164,996 colorectal cancer patients in Germany. PLoS One. 2013 Jul 5;8(7):e68077. doi: 10.1371/journal.pone.0068077. PMID: 23861851; PMCID: PMC3702575.

[11] Abancens M, Bustos V, Harvey H, McBryan J, Harvey BJ. Sexual Dimorphism in Colon Cancer. Front Oncol. 2020 Dec 9;10:607909. doi: 10.3389/fonc.2020.607909. PMID: 33363037; PMCID: PMC7759153.

[12] Majek O, Gondos A, Jansen L, Emrich K, Holleczek B, Katalinic A, Nennecke A, Eberle A, Brenner H; GEKID Cancer Survival Working Group. Sex differences in colorectal cancer survival: population-based analysis of 164,996 colorectal cancer patients in Germany. PLoS One. 2013 Jul 5;8(7):e68077. doi: 10.1371/journal.pone.0068077. PMID: 23861851; PMCID: PMC3702575.

[13] Petrelli F, Tomasello G, Borgonovo K, Ghidini M, Turati L, Dallera P, Passalacqua R, Sgroi G, Barni S. Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer: A Systematic Review and Meta-analysis. JAMA Oncol. 2017 Feb 1;3(2):211-219. doi: 10.1001/jamaoncol.2016.4227. PMID: 27787550.

[14] Hansen IO, Jess P. Possible better long-term survival in left versus right-sided colon cancer - a systematic review. Dan Med J. 2012 Jun;59(6):A4444. PMID: 22677242.

[15] Cai Y, Rattray NJW, Zhang Q, Mironova V, Santos-Neto A, Hsu KS, Rattray Z, Cross JR, Zhang Y, Paty PB, Khan SA, Johnson CH. Sex Differences in Colon Cancer Metabolism Reveal A Novel Subphenotype. Sci Rep. 2020 Mar 17;10(1):4905. doi: 10.1038/s41598-020-61851-0. PMID: 32184446; PMCID: PMC7078199.

[16] Roshanaei G, Safari M, Faradmal J, Abbasi M, Khazaei S. Factors affecting the survival of patients with colorectal cancer using random survival forest. J Gastrointest Cancer. 2020 Nov 10. doi: 10.1007/s12029-020-00544-3. Epub ahead of print. PMID: 33174117.

[17] Xu Y, Ju L, Tong J, Zhou CM, Yang JJ. Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. Sci Rep. 2020 Feb 13;10(1):2519. doi: 10.1038/s41598-020-59115-y. PMID: 32054897; PMCID: PMC7220939.

[18] Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C, Lundin J. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018 Feb 21;8(1):3395. doi: 10.1038/s41598-018-21758-3. PMID: 29467373; PMCID: PMC5821847.

[19] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, Gaiser T, Marx A, Valous NA, Ferber D, Jansen L, Reyes-Aldasoro CC, Zörnig I, Jäger D, Brenner H, Chang-Claude J, Hoffmeister M, Halama N. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. 2019 Jan 24;16(1):e1002730. doi: 10.1371/journal.pmed.1002730. PMID: 30677016; PMCID: PMC6345440.

[20] Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, Wang W, Sheng H, Pu H, Mo H, Zuo Z, Liu Z, Li C, Xie C, Zeng Z, Li W, Hao X, Liu Y, Cao S, Liu W, Gibson S, Zhang K, Xu G, Xu RH. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Sci Transl Med. 2020 Jan 1;12(524):eaax7533. doi: 10.1126/scitranslmed.aax7533. Erratum in: Sci Transl Med. 2020 Apr 22;12(540): PMID: 31894106.

[21] Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019 May 6;9(1):6994. doi: 10.1038/s41598-019-43372-7. PMID: 31061433; PMCID: PMC6502856.

[22] Radespiel-Tröger M, Rabenstein T, Schneider HT, Lausen B. Comparison of tree-based methods for prognostic stratification of survival data. Artif Intell Med. 2003 Jul;28(3):323-41. doi: 10.1016/s0933-3657(03)00060-5. PMID: 12927339.

[23] Dietrich S, Floegel A, Troll M, Kühn T, Rathmann W, Peters A, Sookthai D, von Bergen M, Kaaks R, Adamski J, Prehn C, Boeing H, Schulze MB, Illig T, Pischon T, Knüppel S, Wang-Sattler R, Drogan D. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol. 2016 Oct;45(5):1406-1420. doi: 10.1093/ije/dyw145. Epub 2016 Sep 1. PMID: 27591264.

[24] Rizza S, Copetti M, Rossi C, Cianfarani MA, Zucchelli M, Luzi A, Pecchioli C, Porzio O, Di Cola G, Urbani A, Pellegrini F, Federici M. Metabolomics signature improves the prediction of cardiovascular events in elderly subjects. Atherosclerosis. 2014 Feb;232(2):260-4. doi: 10.1016/j.atherosclerosis.2013.10.029. Epub 2013 Nov 18. PMID: 24468136.

[25] Datema FR, Moya A, Krause P, Bäck T, Willmes L, Langeveld T, Baatenburg de Jong RJ, Blom HM. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. Head Neck. 2012 Jan;34(1):50-8. doi: 10.1002/hed.21698. Epub 2011 Feb 14. PMID: 21322080.

[26] Wang H, Li G. A Selective Review on Random Survival Forests for High Dimensional Data. Quant Biosci. 2017;36(2):85-96. doi: 10.22283/qbs.2017.36.2.85. PMID: 30740388; PMCID: PMC6364686.

[27] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–88.

[28] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics. 2000;28(2).

[29] Ahmed S, Pahwa P, Le D, Chalchal H, Chandra-Kanthan S, Iqbal N, Fields A. Primary Tumor Location and Survival in the General Population With Metastatic Colorectal Cancer. Clin Colorectal Cancer. 2018 Jun;17(2):e201-e206. doi: 10.1016/j.clcc.2017.11.001. Epub 2017 Nov 21. PMID: 29221688.

[30] Wang CB, Shahjehan F, Merchea A, Li Z, Bekaii-Saab TS, Grothey A, Colibaseanu DT, Kasi PM. Impact of Tumor Location and Variables Associated With Overall Survival in Patients With Colorectal Cancer: A Mayo Clinic Colon and Rectal Cancer Registry Study. Front Oncol. 2019 Feb 19;9:76. doi: 10.3389/fonc.2019.00076. PMID: 30838175; PMCID: PMC6389639.

[31] Kamphorst JJ, Cross JR, Fan J, de Stanchina E, Mathew R, White EP, Thompson CB, Rabinowitz JD. Hypoxic and Ras-transformed cells support growth by scavenging unsaturated fatty acids from lysophospholipids. Proc Natl Acad Sci U S A. 2013 May 28;110(22):8882-7. doi: 10.1073/pnas.1307237110. Epub 2013 May 13. PMID: 23671091; PMCID: PMC3670379.

[32] Krall AS, Xu S, Graeber TG, Braas D, Christofk HR. Asparagine promotes cancer cell proliferation through use as an amino acid exchange factor. Nat Commun. 2016 Apr 29;7:11457. doi: 10.1038/ncomms11457. PMID: 27126896; PMCID: PMC4855534.

[33] Knott SRV, Wagenblast E, Khan S, Kim SY, Soto M, Wagner M, Turgeon MO, Fish L, Erard N, Gable AL, Maceli AR, Dickopf S, Papachristou EK, D'Santos CS, Carey LA, Wilkinson JE, Harrell JC, Perou CM, Goodarzi H, Poulogiannis G, Hannon GJ. Asparagine bioavailability governs metastasis in a model of breast cancer. Nature. 2018 Feb 15;554(7692):378-381. doi: 10.1038/nature25465. Epub 2018 Feb 7. Erratum in: Nature. 2018 Apr 4;556(7699):135. PMID: 29414946; PMCID: PMC5898613.

[34] Du F, Chen J, Liu H, Cai Y, Cao T, Han W, Yi X, Qian M, Tian D, Nie Y, Wu K, Fan D, Xia L. SOX12 promotes colorectal cancer cell proliferation and metastasis by regulating asparagine synthesis. Cell Death Dis. 2019 Mar 11;10(3):239. doi: 10.1038/s41419-019-1481-9. PMID: 30858360; PMCID: PMC6412063.

[35]Bos JL. ras oncogenes in human cancer: a review. Cancer Res. 1989 Sep 1;49(17):4682-9. Erratum in: Cancer Res 1990 Feb 15;50(4):1352. PMID: 2547513.

[36] Toda K, Kawada K, Iwamoto M, Inamoto S, Sasazuki T, Shirasawa S, Hasegawa S, Sakai Y. Metabolic Alterations Caused by KRAS Mutations in Colorectal Cancer Contribute to Cell Adaptation to Glutamine Depletion by Upregulation of Asparagine Synthetase. Neoplasia. 2016 Nov;18(11):654-665. doi: 10.1016/j.neo.2016.09.004. Epub 2016 Oct 18. PMID: 27764698; PMCID: PMC5071549.

[37] Li X, Chung ACK, Li S, Wu L, Xu J, Yu J, Wong C, Cai Z. LC-MS-based metabolomics revealed SLC25A22 as an essential regulator of aspartate-derived amino acids and polyamines in KRAS-mutant colorectal cancer. Oncotarget. 2017 Sep 20;8(60):101333-101344. doi: 10.18632/oncotarget.21093. PMID: 29254168; PMCID: PMC5731878.

[38] Knott SRV, Wagenblast E, Khan S, Kim SY, Soto M, Wagner M, Turgeon MO, Fish L, Erard N, Gable AL, Maceli AR, Dickopf S, Papachristou EK, D'Santos CS, Carey LA, Wilkinson JE, Harrell JC, Perou CM, Goodarzi H, Poulogiannis G, Hannon GJ. Asparagine bioavailability governs metastasis in a model of breast cancer. Nature. 2018 Feb 15;554(7692):378-381. doi: 10.1038/nature25465. Epub 2018 Feb 7. Erratum in: Nature. 2018 Apr 4;556(7699):135. PMID: 29414946; PMCID: PMC5898613.

[39] Report for: Asparagine bioavailability governs metastasis in a model of breast cancer. (n.d.). Retrieved April 25, 2021, from https://nature.altmetric.com/details/32797019.

[40] Gallagher, J. (2018, February 07). Food may influence cancer spread. Retrieved April 25, 2021, from https://www.bbc.com/news/health-42976851.

[41] Cai Y, Rattray NJW, Zhang Q, Mironova V, Santos-Neto A, Muca E, Vollmar AKR, Hsu KS, Rattray Z, Cross JR, Zhang Y, Paty PB, Khan SA, Johnson CH. Tumor Tissue-Specific Biomarkers of Colorectal Cancer by Anatomic Location and Stage. Metabolites. 2020 Jun 19;10(6):257. doi: 10.3390/metabo10060257. PMID: 32575361; PMCID: PMC7345993.

[42] Cheng Y, Xie G, Chen T, Qiu Y, Zou X, Zheng M, Tan B, Feng B, Dong T, He P, Zhao L, Zhao A, Xu LX, Zhang Y, Jia W. Distinct urinary metabolic profile of human colorectal cancer. J Proteome Res. 2012 Feb 3;11(2):1354-63. doi: 10.1021/pr201001a. Epub 2011 Dec 28. PMID: 22148915.

[43] Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Chiorean EG, Raftery D. Colorectal cancer detection using targeted serum metabolic profiling. J Proteome Res. 2014 Sep 5;13(9):4120-30. doi: 10.1021/pr500494u. Epub 2014 Aug 15. PMID: 25126899.

[44] Drescher C, Konishi M, Ebner N, Springer J. Loss of muscle mass: current developments in cachexia and sarcopenia focused on biomarkers and treatment. J Cachexia Sarcopenia Muscle. 2015 Dec;6(4):303-11. doi: 10.1002/jcsm.12082. Epub 2015 Nov 18. PMID: 26676067; PMCID: PMC4670737.

[45] Johnson CH, Spilker ME, Goetz L, Peterson SN, Siuzdak G. Metabolite and Microbiome Interplay in Cancer Immunotherapy. Cancer Res. 2016 Nov 1;76(21):6146-6152. doi: 10.1158/0008-5472.CAN-16-0309. Epub 2016 Oct 11. PMID: 27729325; PMCID: PMC5093024.

[46] Dai ZL, Li XL, Xi PB, Zhang J, Wu G, Zhu WY. L-Glutamine regulates amino acid utilization by intestinal bacteria. Amino Acids. 2013 Sep;45(3):501-12. doi: 10.1007/s00726-012-1264-4. Epub 2012 Mar 24. PMID: 22451274.

[47] Penney KL, Tyekucheva S, Rosenthal J, El Fandy H, Carelli R, Borgstein S, Zadra G, Fanelli GN, Stefanizzi L, Giunchi F, Pomerantz M, Peisch S, Coulson H, Lis R, Kibel AS, Fiorentino M, Umeton R, Loda M. Metabolomics of Prostate Cancer Gleason Score in Tumor Tissue and Serum. Mol Cancer Res. 2021 Mar;19(3):475-484. doi: 10.1158/1541-7786.MCR-20-0548. Epub 2020 Nov 9. PMID: 33168599.

[48] Karaglani M, Gourlia K, Tsamardinos I, Chatzaki E. Accurate Blood-Based Diagnostic Biosignatures for Alzheimer's Disease via Automated Machine Learning. J Clin Med. 2020 Sep 18;9(9):3016. doi: 10.3390/jcm9093016. PMID: 32962113; PMCID: PMC7563988.

[49] Orlenko A, Kofink D, Lyytikäinen LP, Nikus K, Mishra P, Kuukasjärvi P, Karhunen PJ, Kähönen M, Laurikka JO, Lehtimäki T, Asselbergs FW, Moore JH. Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. Bioinformatics. 2020 Mar 1;36(6):1772-1778. doi: 10.1093/bioinformatics/btz796. PMID: 31702773; PMCID: PMC7703753.

[50] Surveillance Guidelines for Measles, Rubella and Congenital Rubella Syndrome in the WHO European Region. Geneva: World Health Organization; 2012 Dec. Annex 3, Collection, storage and shipment of specimens for laboratory diagnosis and interpretation of results. Available from: https://www.ncbi.nlm.nih.gov/books/NBK143256/

[51] Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Quality of reporting of cancer prognostic marker studies: association with reported prognostic effect. J Natl Cancer Inst. 2007 Feb 7;99(3):236-43. doi: 10.1093/jnci/djk032. PMID: 17284718.

[52] Hemingway H, Philipson P, Chen R, Fitzpatrick NK, Damant J, Shipley M, Abrams KR, Moreno S, McAllister KS, Palmer S, Kaski JC, Timmis AD, Hingorani AD. Evaluating the quality of research into a single prognostic biomarker: a systematic review and meta-analysis of 83 studies of C-reactive protein in stable coronary artery disease. PLoS Med. 2010 Jun 1;7(6):e1000286. doi: 10.1371/journal.pmed.1000286. PMID: 20532236; PMCID: PMC2879408.

[53] Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011 May;12(3):189-202. doi: 10.1093/bib/bbq073. Epub 2011 Feb 7. PMID: 21300697; PMCID: PMC3088312.

[54] Ioannidis JPA, Bossuyt PMM. Waste, Leaks, and Failures in the Biomarker Pipeline. Clin Chem. 2017 May;63(5):963-972. doi: 10.1373/clinchem.2016.254649. Epub 2017 Mar 7. PMID: 28270433.

Appendix

Supplementary Table 1. Demographic characteristics and clinical factors for each subgroup. RCC = right-sided colon cancer, LCC = left-sided colon cancer.

Subgroup	Stage I (n=47)		Stage II (n=86)		Stage III (n=64)	
	RCC (n=22)	LCC (n=25)	RCC (n=44)	LCC (n=42)	RCC (n=32)	LCC (n=32)
Sex, n						
Males	10	15	23	25	15	14
Females	12	10	21	17	17	18
Age, mean (SD)						
Males	73.9 (6.5)	69.3 (5.8)	72.9 (7.8)	72.2 (8.5)	73.5 (7.8)	63.7 (5.8)
Females	72.1 (6.2)	69.6 (7.6)	73.5 (9.8)	69.1 (7.8)	72.2 (6.6)	71.1 (6.0)
5-year Overall survival rate, %[a]						
Males	87.5	85.1	76.5	74.1	47.1	78.6
Females	100	100	82.9	100	65.2	61.8
5-year Recurrence-free survival rate, %[a]						
Males	90.0	78.3	86.8	67.6	82.1	70.1
Females	90.9	100.0	87.5	87.4	76.0	68.8

[a] The survival rates were calculated using the Kaplan-Meier estimation method.

Supplementary Table 2. Prognosis among different subgroups of patients after combining stage I and II together to allow analysis of Anatomic location: RCC vs. LCC females

5-year Overall Survival (OS)								
Subgroup	**RCC males**		**RCC females**		**LCC males**		**LCC females**	
Event[c]	0	1	0	1	0	1	0	1
Stage a[a]	28	5	30	3	32	8	27	0
Stage b[b]	8	7	12	5	11	3	12	6
Total	36	12	42	8	43	11	39	6
5-year Recurrence-free Survival (RFS)								
Subgroup	**RCC males**		**RCC females**		**LCC males**		**LCC females**	
Event[d]	0	1	0	1	0	1	0	1
Stage a[a]	30	3	30	3	32	8	25	2
Stage b[b]	13	2	14	3	10	4	13	5
Total	43	5	44	6	42	12	38	7

[a] Stage a combines Stage I and II together which refers to earlier stages.
[b] Stage b refers to Stage III.
[c] Event = death. 1= the event occurred within 5 years of follow-up, 0 = the event did not occur within 5 years of follow-up (survived or censored).
[d] Event = recurrence. 1= recurrence occurred within 5 years of follow-up, 0 = recurrence does not occur within 5 years of follow-up. If a patient died before CRC recurrence, it will be counted towards a death event.

Notes:
If a patient died for any reason without recurrence within follow-up time (no more than 5 years), it was only counted toward a death event in OS.
If a patient experienced recurrence but did not die within follow-up time (no more than 5 years), it was only counted toward a recurrence event in RFS.
If a patient experienced recurrence and then died within follow-up time (no more than 5 years), it was counted toward a recurrence event in RFS and a death event in OS.
If a patient experienced neither recurrence nor death within follow-up time (no more than 5 years), the event was 0 for both OS and RFS.
Due to the absence of death events among females at clinical stage I, we regarded stage I and II patients as “stage a” (early stage) and stage III patients as “stage b” (late stage). Still, all the LCC females

Supplementary Table 3. Associations between individual metabolites and 5-year overall survival (OS) by sex, adjusted for anatomic location, clinical stages, and age).

Metabolite name	Females			Males			Int. Sex *P* value *
	HR	95% CI	*P* value[a]	HR	95% CI	*P* value[a]	
Acetyl-lysine	0.96	0.74 - 1.25	0.786	0.83	0.72 - 0.96	**0.012**	0.342
Adenosine	0.91	0.70 - 1.19	0.507	1.29	1.03 - 1.62	**0.026**	**0.044**
Alanine	1.05	0.77 - 1.42	0.762	0.77	0.61 - 0.98	**0.034**	0.096
Argininosuccinic acid	0.93	0.71 - 1.23	0.613	0.74	0.58 - 0.93	**0.010**	0.154
Asparagine	1.45	0.87 - 2.42	0.154	0.72	0.55 - 0.96	**0.025**	**0.029**
Carnitine	0.62	0.04 - 9.47	0.733	0.56	0.34 - 0.93	**0.026**	0.881
Citrulline	1.66	0.98 - 2.81	0.061	0.65	0.46 - 0.92	**0.014**	**0.002**
Glycerol 3-phosphate	3.64	1.30 - 10.2	**0.014**	0.91	0.47 - 1.77	0.777	**0.017**
Hypoxanthine	1.04	0.35 - 3.13	0.943	0.65	0.44 - 0.95	**0.027**	0.444
LysoPC(16:0)	1.54	1.04 - 2.27	**0.031**	0.85	0.66 - 1.11	0.244	**0.008**
Ornithine	0.96	0.56 - 1.66	0.895	0.68	0.47 - 0.97	**0.035**	0.316
Serine	1.24	0.65 - 2.38	0.519	0.55	0.37 - 0.81	**0.002**	**0.035**
Spermine	1.40	1.01 - 1.93	**0.041**	1.03	0.83 - 1.27	0.813	0.086
Succinate	0.34	0.12 - 0.96	**0.042**	1.77	1.21 - 2.58	**0.003**	**0.004**
Threonine	1.11	0.67 - 1.86	0.685	0.61	0.44 - 0.85	**0.004**	**0.035**
UDP-D-Glucose	0.81	0.67 - 0.97	**0.023**	1.15	0.95 - 1.40	0.161	**0.012**
Uracil	1.21	0.54 - 2.69	0.643	0.44	0.28 - 0.70	**0.001**	**0.024**
Xanthosine	1.21	0.86 - 1.71	0.283	0.71	0.54 - 0.94	**0.016**	**0.027**

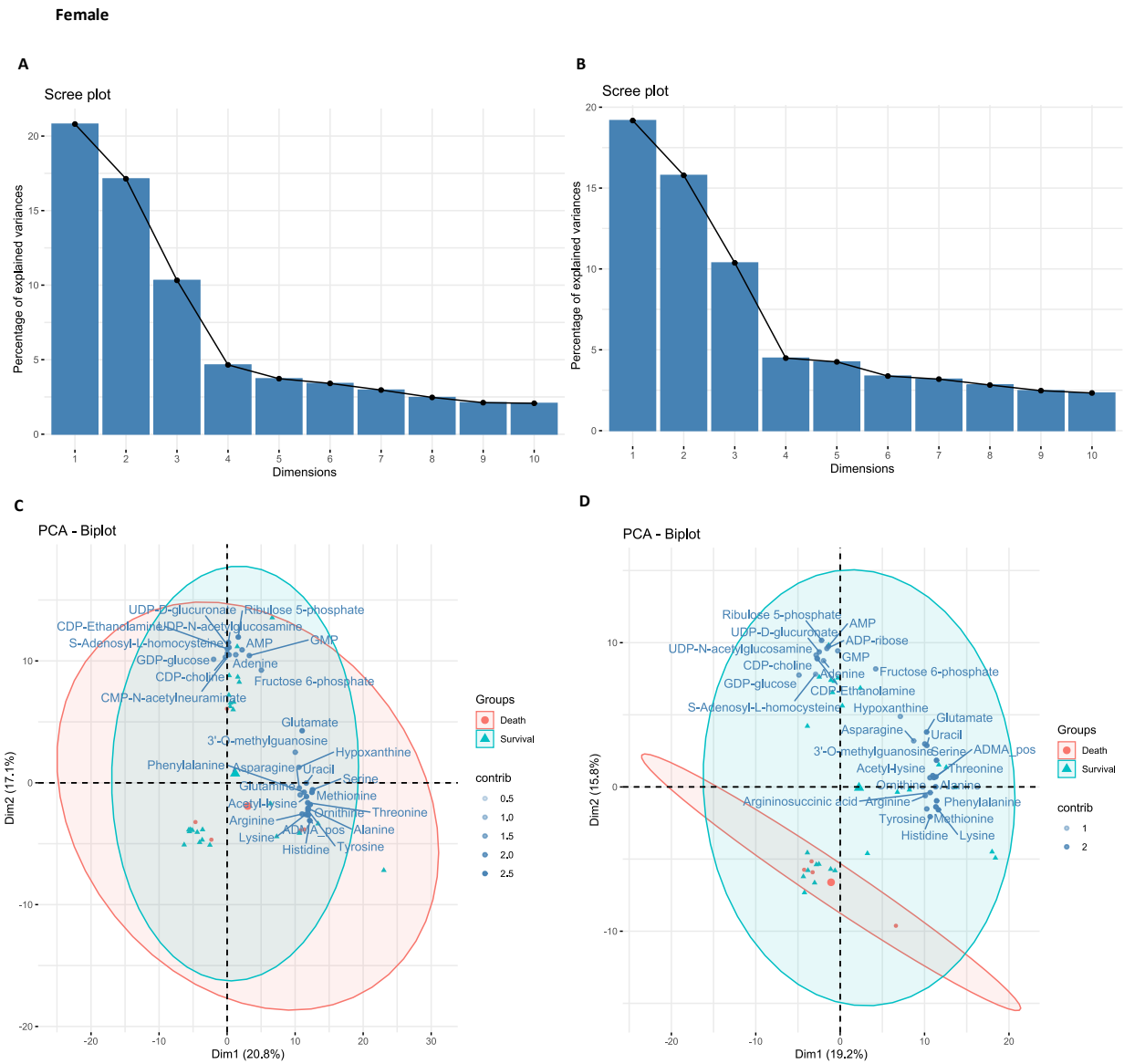
[a] Raw *P* value before FDR adjustment. *Each metabolite with sex-interaction *P* value. The abundance of each metabolite was treated as a continuous variable and was $\log_2$ transformed.

Supplementary Table 4. Associations between individual metabolites and 5-year recurrence-free survival (RFS) by sex, adjusted for anatomic location, clinical stages, and age).

Metabolite name	Females			Males			Int. Sex *P* value *
	HR	95% CI	*P* value[a]	HR	95% CI	*P* value[a]	
Acetyl-lysine	0.97	0.75 - 1.25	0.795	0.79	0.68 - 0.93	**0.003**	0.102
Alanine	1.09	0.79 - 1.51	0.607	0.75	0.59 - 0.95	**0.017**	0.070
AMP	1.1	0.68 - 1.78	0.699	0.71	0.53 - 0.95	**0.019**	0.113
Argininosuccinic acid	1.02	0.77 - 1.34	0.906	0.71	0.54 - 0.93	**0.012**	**0.041**
Asparagine	1.45	0.91 - 2.32	0.119	0.74	0.56 - 0.98	**0.039**	**0.009**
Carnitine	3.23	0.23 - 44.44	0.382	0.38	0.23 - 0.65	**<0.001**	0.115
CMP	0.97	0.56 - 1.67	0.908	0.70	0.51 - 0.96	**0.028**	0.278
Creatinine	0.72	0.41 - 1.26	0.251	1.63	1.13 - 2.36	**0.009**	**0.033**
Cytidine	0.56	0.34 - 0.9	**0.017**	0.75	0.49 - 1.15	0.185	0.246
Fructose 6-phosphate	0.82	0.53 - 1.26	0.355	0.68	0.49 - 0.95	**0.025**	0.556
Glutathione	0.74	0.57 - 0.95	**0.018**	0.92	0.80 - 1.06	0.236	0.104
Glutathione disulfide	0.73	0.58 - 0.91	**0.006**	0.82	0.68 - 0.99	**0.037**	0.329
GMP	1.05	0.69 - 1.61	0.818	0.74	0.57 - 0.96	**0.024**	0.157
Hypoxanthine	1.87	0.57 - 6.19	0.304	0.32	0.21 - 0.51	**<0.001**	**0.009**
LysoPC(16:1)	1.05	0.75 - 1.47	0.765	0.77	0.61 - 0.96	**0.023**	0.100
LysoPE(18:1)	0.98	0.73 - 1.31	0.897	0.83	0.69 – 1.00	**0.049**	0.294
LysoPE(20:1)	0.97	0.77 - 1.23	0.812	0.84	0.71 - 0.98	**0.028**	0.218
LysoPE(22:5)	1.08	0.70 - 1.67	0.721	0.70	0.53 - 0.91	**0.009**	0.070
LysoPE(18:2)	1.07	0.70 - 1.63	0.767	0.73	0.55 - 0.95	**0.021**	0.119
Serine	1.39	0.72 - 2.69	0.329	0.59	0.41 - 0.85	**0.005**	**0.013**
Sphinganine-1-phosphate	0.92	0.59 - 1.44	0.715	0.67	0.48 - 0.94	**0.022**	0.206
Stearamide	0.91	0.62 - 1.33	0.629	0.70	0.53 - 0.92	**0.011**	0.206
Threonine	0.96	0.58 - 1.60	0.885	0.65	0.47 - 0.90	**0.009**	0.086
Xanthine	0.59	0.36 - 0.98	**0.043**	0.69	0.49 - 0.98	**0.036**	0.869
Xanthosine	1.01	0.73 - 1.38	0.972	0.72	0.52 - 0.99	**0.044**	0.112

[a] Raw *P* value before FDR adjustment. *Each metabolite with sex-interaction *P* value. The abundance of each metabolite was treated as a continuous variable and was log2 transformed.

Principal Component Analysis results for females and males



Supplementary Figure 1. Results of Principal Component Analysis (PCA) by sex. Screen plot for females (A) and for males (B) illustrated the explained variance by top 10 components. PCA biplot for females (C) and males (D) indicated a poor overall survival (OS) classification ability by considering the first two components based on metabolite abundance. Amino acids (e.g., threonine, alanine, serine, and asparagine etc.) were successfully distinguished and clustered around at the positive direction of dimension 2 for both females (C) and males (D).

Supplementary Table 5. Eigenvalue and variance of components retained for females and males. We only analyzed components with an eigenvalue greater than 1 based on the Kaiser rule[1]. For females, 19 dimensions were adopted in further analysis, which accounted for 83.23% of the total variance. For males, 20 dimensions were included with a cumulative variance of 83.70%.

	Females			Males		
	Eigenvalue	**Variance percent**	**Cumulative variance percent**	**Eigenvalue**	**Variance percent**	**Cumulative variance percent**
Dim.1	18.93	20.81	20.81	17.45	19.18	19.18
Dim.2	15.59	17.13	37.94	14.37	15.79	34.96
Dim.3	9.39	10.32	48.26	9.44	10.38	45.34
Dim.4	4.23	4.65	52.91	4.08	4.49	49.83
Dim.5	3.39	3.73	56.64	3.87	4.25	54.08
Dim.6	3.10	3.41	60.05	3.07	3.38	57.46
Dim.7	2.70	2.96	63.01	2.90	3.19	60.65
Dim.8	2.25	2.47	65.48	2.57	2.82	63.47
Dim.9	1.93	2.12	67.60	2.25	2.48	65.95
Dim.10	1.89	2.08	69.67	2.12	2.33	68.28
Dim.11	1.73	1.90	71.58	1.77	1.95	70.23
Dim.12	1.68	1.85	73.43	1.69	1.86	72.08
Dim.13	1.49	1.64	75.06	1.63	1.79	73.87
Dim.14	1.41	1.55	76.62	1.52	1.67	75.54
Dim.15	1.39	1.53	78.15	1.47	1.61	77.15
Dim.16	1.29	1.41	79.56	1.42	1.56	78.72
Dim.17	1.22	1.34	80.90	1.28	1.41	80.13
Dim.18	1.11	1.22	82.12	1.16	1.27	81.40
Dim.19	1.01	1.11	83.23	1.08	1.19	82.58
Dim.20	-	-	-	1.01	1.11	83.70

[1] KAISER, H. E (1960). The application of electronic computers to factor analysis. Education & Psychological Measurement, 20, 14I-151.

Supplementary Table 6. Statistically significant results using CPH analysis for individual top 19 components for females and top 20 components (with eigenvalues > 1) for males adjusted for clinical stage, tumor location, and age. No component was associated with OS for female patients. Component 5 and 19 were associated with RSF among females. Metabolites noted with (+) were positively associated with the corresponding component, and the higher the absolute value of loading, the greater the relationship, vice versa. A component with HR > 1 was associated with an increased risk of poorer prognosis. E.g., component 5 was associated with both poor OS and RSF for males. Males with higher levels of succinate, creatinine, L-phenylalanine, and lower levels of carnitine and cytidine have higher risks of all-cause mortality and CRC recurrence.

Outcome	Component number	Influential metabolites: (+): loading > 0.2; (-): loading <- 0.2	HR (95% CI)	P	C-index (Se)	Any variable Violated PH assumption
OS (Females)	None	-	-	-	-	-
OS (Males)	5	Succinate (+), Creatinine (+), L-Phenylalanine (-), Carnitine (-), Cytidine (-)	1.20 (1.02-1.42)	0.026	0.751 (0.052)	No
RSF (Females)	5	Palmitic acid (+), Diacetylspermine (+), Stearic acid (+), Oleic acid (+), Glutathione disulfide (-), Adenosine (-), Xanthine (-)	1.72 (1.22-2.43)	0.002	0.799 (0.048)	No
	19	Vitamin E (+), CMP, PC(36:2) (+), Lactate (+), D-Glucuronate (+), Adenosine (+),	1.91 (1.09-3.34)	0.023	0.749 (0.063)	No
RSF (Males)	2	Ribulose 5-phosphate (+), AMP (+), ADP-ribose (+), L-Phenylalanine (-), Carnitine (-), Cytidine (-)	0.87 (0.76-0.98)	0.027	0.716 (0.06)	Yes
	5	Succinate (+), Creatinine (+), L-Phenylalanine (-), Carnitine (-), Cytidine (-)	1.34 (1.13-1.59)	0.001	0.706 (0.07)	Yes