随机生存森林在结直肠癌患者基因数据预后分析中的应用研究*

山东第二医科大学公共卫生学院(261053)

穆华夏　卜伟晓　高梦瑶　苏维强　韩　梅　徐雅琪　陶子琨　杨　希　石福艳　王清华　孔雨佳△　王素珍△

【提　要】 **目的**　应用随机生存森林模型探讨基因数据中结直肠癌患者预后影响因素。**方法**　利用 TCGA 数据库中结直肠癌基因表达数据,对差异表达基因进行筛选,结合临床与生存信息构建 RSF 模型,并与传统 Lasso-Cox 回归模型进行比较。**结果**　通过 RSF 模型得到包括 *HAND*1(VIMP=0.090)和 *PCOLCE*2(VIMP=0.075)基因表达在内的 13 个影响结直肠癌患者预后的重要因素,并分析了病理学 N 分期、*PCOLCE*2 基因及 *IGSF*9 基因变量之间的交互作用。与 Lasso-Cox 模型比较结果显示,尽管 RSF 模型预测错误率略高(1-*C*-index:训练集:0.296 *vs.* 0.213;测试集:0.369 *vs.* 0.332),但具有更好的模型校准度(IBS:训练集:0.205 *vs.* 0.214;测试集:0.210 *vs.* 0.221)。**结论**　RSF 模型在处理右删失生存数据的分析时具有良好的表现,能发现重要的影响因素以及变量之间的交互作用,为结直肠癌患者预后状况的改善和生命质量的提升提供了科学依据。

【关键词】 随机生存森林　Lasso-Cox 回归　结直肠癌　基因数据　预后分析

【中图分类号】 R735.3　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.04.011

Application of Random Survival Forest in Prognosis Analysis of Genetic Data in Patients with Colorectal Cancer

Mu Huaxia, Bu Weixiao, Gao Mengyao, et al (*School of Public Health*, *Shandong Second Medical University*, *Weifang* 261053)

【**Abstract**】 **Objective**　To explore the prognostic factors of colorectal cancer patients in gene data using random survival forest model. **Method**　The differentially expressed genes were screened using the gene expression data of colorectal cancer in TCGA database, and combined with clinical and survival information. The RSF model is constructed and compared with the traditional Lasso-Cox regression model. **Results**　The RSF model obtained 13 important factors affecting the prognosis of colorectal cancer patients, including *HAND*1 (VIMP = 0.090) and *PCOLCE*2 (VIMP = 0.075) genes, and analyzed the interaction between pathological N, *PCOLCE*2 gene and *IGSF*9 gene variables. Compared with Lasso-Cox model, the RSF model has better model calibration (IBS: training set: 0.205 *vs.* 0.214; test set: 0.210 *vs.* 0.221) although its prediction error rate is slightly higher (1-*C*-index: training set: 0.296 *vs.* 0.213; test set: 0.369 *vs.* 0.332). **Conclusion**　RSF model has a good performance in processing the analysis of right censored survival data, can find important influencing factors and the interaction between variables, and provide scientific basis for the improvement of prognosis and quality of life of colorectal cancer patients.

【**Key words**】 Random survival forest; Lasso-Cox regression; Colorectal cancer; Genetic data; Prognostic analysis

结直肠癌(Colorectal cancer, CRC)是最常见的消化系统恶性肿瘤,近年来发病率和死亡率持续上升,已成为我国发病率第三、死亡率第五的癌症[1],给公共卫生资源和患者家庭带来沉重的疾病负担。由于目前结直肠癌的主要致病机制尚无明确定论,尽管可以通过手术和化疗的方法治疗,术后仍存在相当比例的患者出现癌细胞的转移和复发,预后效果仍不理想。因此,可靠的结直肠癌预后分子标记物的确定及与免疫细胞的关联性研究对于改善结直肠癌患者的预后有重要的临床价值。

近年来,测序与基因组学数据采集技术的发展产生了大量的高维或超高维的数据,这对统计分析方法提出了更高的要求。传统生存数据通常基于 Cox 回归构建的预后模型,然而在处理高维数据,尤其是数据维度远大于样本量时,会出现变量之间多重共线性导致模型失真的结果。Lasso-Cox 模型通过在 Cox 回归引入 L1 正则化惩罚函数实现对特征变量的选择和压缩以防止模型过度拟合[2-3]。但是,传统生存分析和 Lasso-Cox 均需要资料满足比例风险(proportional hazards, PH)假定[4],在实际使用中有很大局限性。Ishwaran 等提出的随机生存森林(random survival forest, RSF)模型,将机器学习方法和传统生存分析相结合,克服传统生存分析方法需要资料满足 PH 假定的限制[5]。

本研究分别基于 RSF 模型和 Lasso-Cox 回归模型两种方法对结直肠癌基因数据进行挖掘,探讨其在结直肠癌患者预后研究中的效果,同时为临床提供有助于筛选高危人群和改进治疗设计的潜在生物学标记物,为预测患者死亡风险以及提供针对性的干预措施等降低患者死亡风险提供科学依据。

*基金项目:国家自然科学基金(82003560);山东省自然科学基金(ZR2020MH340, ZR2023MH313);山东省教育厅教改项目(M2021174,M2021327)

△通信作者:孔雨佳,E-mail: yujia_kyj80@163.com;王素珍,E-mail: wangsz@wfmc.edu.cn

资料与方法

1.数据来源

从基因组数据共享数据库(genomic data commons, GDC)下载癌症基因组图谱(the cancer genome atlas, TCGA)中结肠癌(colon adenocarcinoma, COAD)和直肠癌(rectal adenocarcinoma, READ)数据集的基因表达量数据、表型数据以及生存数据。构建基因表达矩阵,筛选全部癌症样本的基因表达数据,并将筛选后的基因数据合并生存数据及表型数据,使用“rsample”包将数据按1∶1的比例随机划分训练集和测试集,分别用于预后模型的建立和验证。本研究共获得病例样本689例,组学信息20107个。由于TCGA数据库中有来源于同一样本的不同组织数据,经整理去重后最终纳入606名患者的信息进行预后分析。对非单调高维缺失数据采用“missForest”包基于随机森林(random forests, RF)算法的机器学习方法进行插补[6]。

2.差异表达基因的筛选

对数据集的基因表达数据进行预处理标准化并构建比较矩阵,使用“limma”软件包识别结直肠癌组织与癌旁组织的差异表达基因(differential expression gene, DEGs)。选取|log2FC|>2且调整后的$P<0.05$的基因纳入研究。用“EnhancedVolcano”包绘制火山图。

3.预后预测模型的构建

为评价差异表达基因与结直肠癌患者预后的相关性,利用“survival”包对TCGA队列中每个DEGs进行单因素Cox回归分析,评估其与CRC患者预后的相关性。以$P<0.05$筛选与患者预后相关的基因。

(1)稀疏的Cox模型

通过“glment”包构建Lasso-Cox回归模型以缩小预后相关基因的范围,用cv.glmnet()函数进行10折交叉验证选择最小λ值,保留关键基因及其系数,利用关键基因构建预后模型[7]。

(2)随机生存森林模型

随机生存森林模型是基于生存树的非参数非线性的集成机器学习方法[8],是RF基于右删失生存数据的扩展。模型应用boot-strap和随机节点分裂来生长独立二元生存树,将所有的树集成形成RSF。由于RSF选取数据的大量特征作为分裂节点来构建模型,保留了冗杂的变量,可通过变量筛选了解各变量在建模时的作用。VIMP(variable importance)法和最小深度法(minimal depth, MD)是RSF模型中常用的变量筛选方法,本研究采用VIMP法作为主要方法[9]。此外,RSF能够实现自动拟合交互作用和控制过度拟合[5],帮助识别疾病预后的共同影响因素。

4.构建RSF模型的基本流程

(1)构建样本分析集 通过Bootstrap有放回的随机抽取63%的原始数据,其余的37%袋外数据(out-of-bag data, OOB)可以通过计算累积风险函数(cumulative hazard function, CHF)获取预测的错误率,有利于降低集合树的泛化误差[10]。

(2)构建对应的生存树 在树的每个节点随机选择的mtry个候选变量中选择使子节点间生存差异最大的作为节点进行分裂,当子节点的样本数目小于预先设数目时停止分裂[11]。通过对数秩(log-rank)或对数秩得分(log-rank score)的分裂准则方法[12]比较不同的生存曲线,以评价分裂变量及分裂点的有效性。

(3)估计RSF模型的累积风险 对单棵树在节点处基于Nelson-Aalen估计量计算集成累积风险函数。利用“randomSurvivalForest”包进行模型参数调优及RSF模型构建。

5.预后模型的评价

采用“randomSurvivalForest”包中自带的程序及函数C-index()计算Harrell一致性指数(concordance index, C-index),衡量生存模型的预测准确性[13]。本研究用1-C-index表示预测错误率,预测错误率在0到1之间,0为预测效果最好。采用“pec”包计算时间依赖的Brier分数(integrated brier score, IBS)以评估生存分析的模型校准度。

6.软件实现

本研究模型的建立与评价均采用R 4.2.2,取检验水准$\alpha=0.05$。

结 果

1.数据预处理

数据主要分为基因数据和临床数据,基因数据无缺失,临床数据存在多变量、任意模式的缺失。采用RF算法对缺失值插补(图1)。

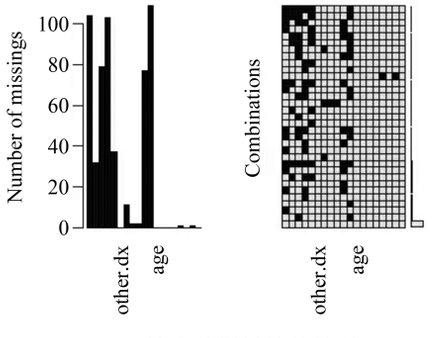

图1 缺失数据插补前情况

2.DEGs的筛选

从TCGA获取的635个CRC癌组织和51个癌旁组织基因的mRNA表达矩阵中,共筛选出2421个

mRNA 差异表达基因，其中表达上调的基因 2170 个，下调的基因 251 个（图 2）。

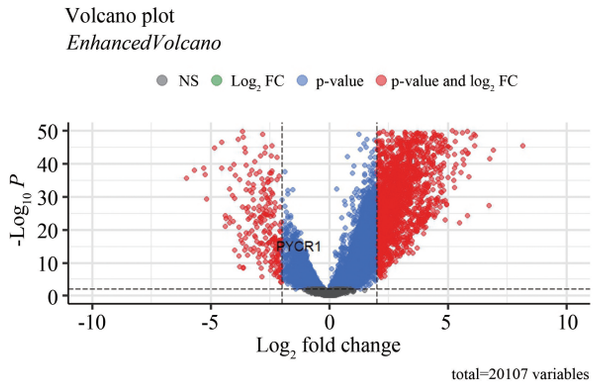

图 2　火山图

对 2421 个 DEGs 进行单因素 Cox 回归分析得到 235 个 CRC 预后相关基因（$P<0.05$）。筛选出来的预后相关基因与插补后的临床资料合并，最终的数据集有 606 个观测，254 个变量，按照 1∶1 的比例划分得到训练集和测试集。

3.基因数据模型构建

（1）Lasso-Cox 模型

将单因素 Cox 回归中有统计学意义的 235 个 DEGs 及 19 个临床变量通过 10 折交叉验证选择最优参数 λ（λ.min＝0.05892）进行模型构建，得到包含 10 个基因及 3 个临床特征预后模型（图 3）。

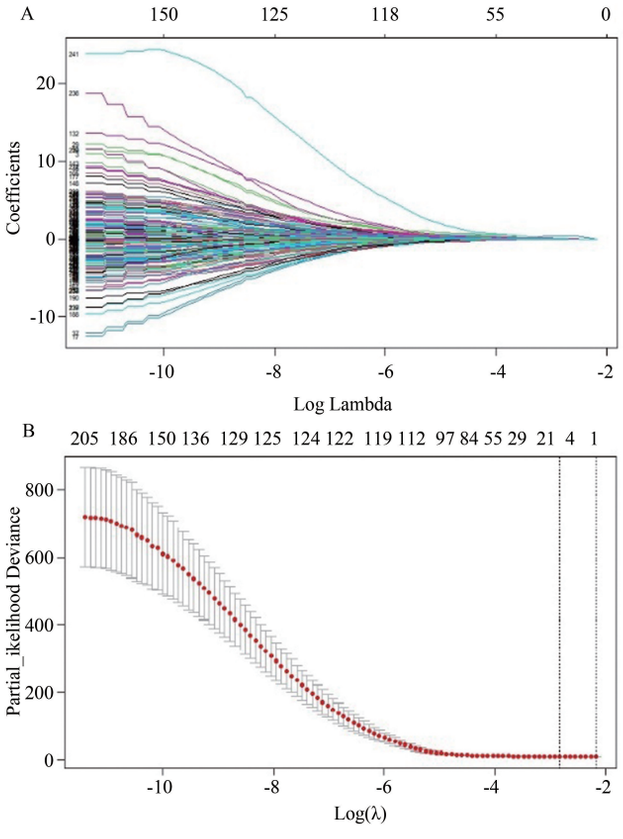

图 3　基于 Lasso-Cox 回归的变量筛选（A）及交叉验证（B）构建模型

Lasso 回归筛选非零系数的变量，将筛选出的 13 个变量根据系数从大到小排列如表 1 所示。

表 1　Lasso 回归筛选的变量

变量名称	系数
病理学 N 分期	0.394632864
年龄	0.060794807
PCOLCE2	0.040253926
CDK5R2	0.035507848
C8G	0.03143241
he 淋巴结阳性计数	0.011252171
IGSF9	0.008640205
HAND1	0.003012272
SFTA2	0.001998719
SCGB2A1	−0.005060272
TRPA1	−0.011174303
IGHV1.45	−0.012156465
HEPACAM2	−0.016763393

（2）RSF 模型

构建 RSF 模型，前 4 个重要的分裂节点分别是 he 淋巴结阳性计数、*CDK5R2*、*PCOLCE2* 、*HAND1*（图 4）。

对 RSF 模型进行参数调优，当最优参数 nodesize＝3 与 mtry＝67 时，训练集模型的性能最优（图 5-A）。该模型通过在 log-rank 分裂规则下选择合适的 ntree 参数构建随机生存森林模型（图 5-B）。模型共生成 500 个二元生存树，平均每个生存树有 3 个终端节点，基于 OOB 验证预测模型生存结局的错误率为 29.57%。该随机生存森林模型在测试集中进行预测的错误率为 36.91% 。

对 13 个自变量的重要性由强到弱进行排序（图 6）。

对筛选出的 13 个变量进行交互作用分析，部分结果如表 2 所示。

由表 2 可知，变量间交互作用最大的是病理学 N 分期和 *PCOLCE2* 基因，其次是 *PCOLCE2* 基因和 *IGSF9* 基因。绘制变量间交互作用对生存时间影响的 coplot 图（图 7）。结果显示，在 *IGSF9* 表达量不变的情况下，*PCOLCE2* 基因表达量和病理学 N 分期的交互作用使生存时间降低。

4.模型比较

（1）预测错误率

计算训练集和测试集的模型一致性错误率。在训练集和测试集中 RSF 模型一致性错误率均高于 Lasso-Cox 模型（训练集：0.296 *vs.* 0.213；测试集：0.369 *vs.* 0.332）。

(2)预测误差曲线

基于 100 次 bootstrap 交叉验证计算两种模型的综合 Brier 分数。训练集中 RSF 模型的综合 Brier 分数为 0.205, Lasso-Cox 模型的综合 Brier 分数为 0.214;测试集中 RSF 模型综合 Brier 分数为 0.210, Lasso-Cox 模型的综合 Brier 分数为 0.221。两模型均有较好的校准度,且 RSF 模型较 Lasso-Cox 模型的校准度更高。绘制预测误差曲线如下图 8 所示。

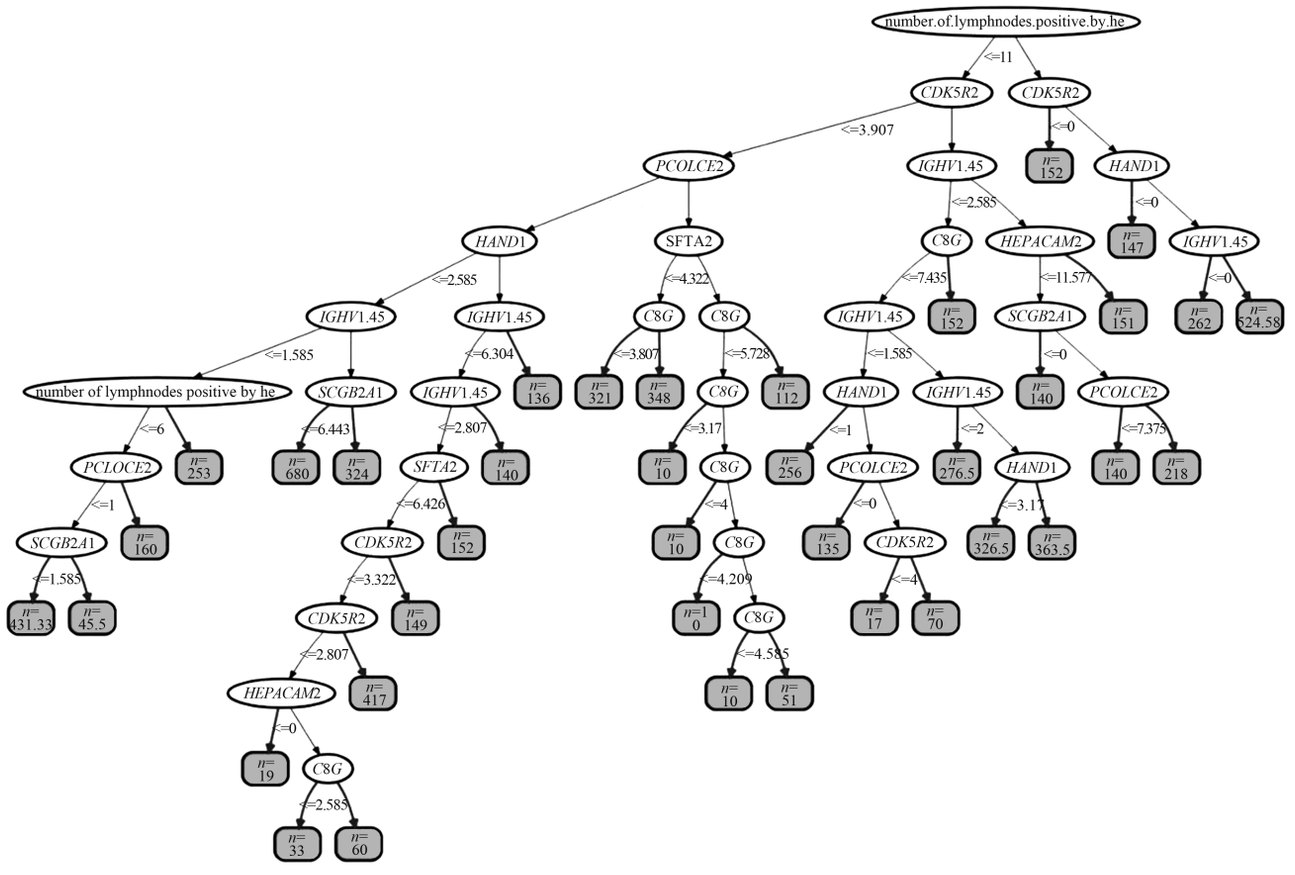

图 4 RSF 树模型

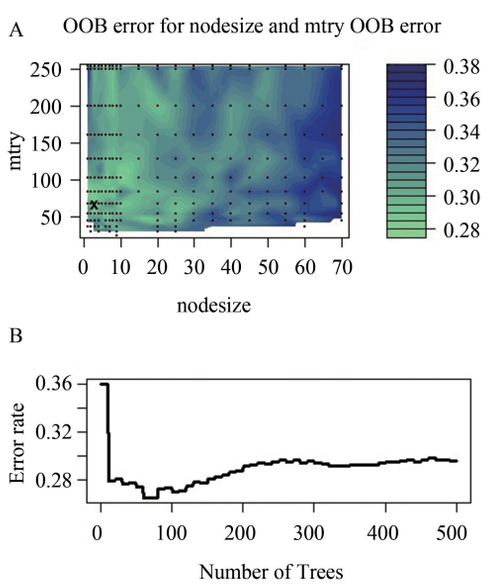

图 5 RSF 模型 nodesize 与 mtry 参数调整(A)及 ntree 参数选择(B)

通过预测误差曲线可知 RSF 模型和 Lasso-Cox 模型整体的预测错误率与参考值基本持平,说明两种生存模型预测效果与真实值较为接近。在 1500 天之前,两种模型的预测误差基本持平,1500 天之后,RSF 模型的预测误差较 Lasso-Cox 更低。综上,可以认为 RSF 模型在生存分析中的预测误差持平甚至小于 Lasso-Cox 模型,是一种较为准确的生存分析方法。

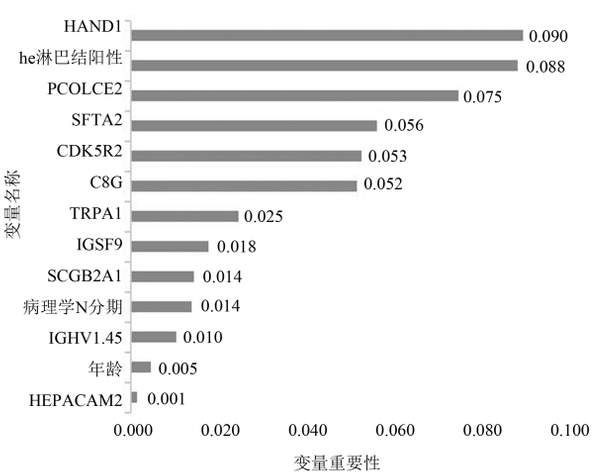

图 6 各变量重要性排序

表 2　变量间交互作用

	Paired	Additive	Difference
PCOLCE2：pathologic N	0.026970174	0.02163427	0.005335905
IGSF9：*PCOLCE2*	0.022227125	0.017180808	0.005046317
SFTA2：number of lymphnodes positive by he	0.04813595	0.043938915	0.004197034
HAND1：number of lymphnodes positive by he	0.041680583	0.037840656	0.003839927
TRPA1：age	0.010420548	0.00682445	0.003596098
HEPACAM2：*IGSF9*	0.004638575	0.001107956	0.00353062
IGSF9：number of lymphnodes positive by he	0.029609765	0.02632637	0.003283396
HEPACAM2：number of lymphnodes positive by he	0.027621678	0.024695563	0.002926115
HEPACAM2：*IGHV1.45*	0.006793119	0.00402776	0.00276536
CDK5R2：*SCGB2A1*	0.012958679	0.010373447	0.002585232
IGSF9：pathologic N	0.009966406	0.00754092	0.002425486

＊：paired 表示可能存在交互作用两变量的 VIMP 值；Additive 是两量单独作用的重要性之和；Derference 是 paired 与 Additire 之差，即反映两变量交互作用大小。

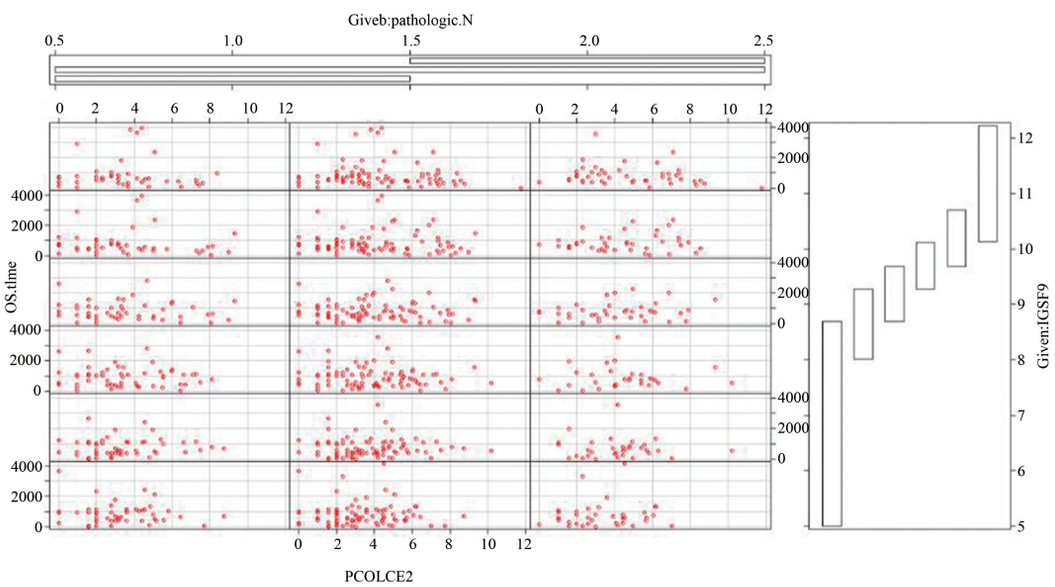

图 7　变量的交互作用对生存时间的影响

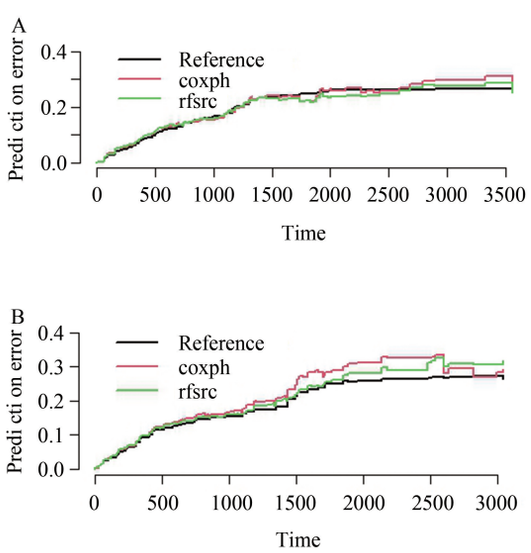

图 8　训练集（A）和测试集（B）的预测误差曲线比较

讨　论

信息技术和基因测序技术的不断发展催生了大量组学高维数据。Lasso-Cox 模型作为一种有效的处理高维生存数据的方法得到了较为广泛的应用，常用来识别预后相关因素，可利用预后指数对患者进行预后分组。RSF 模型的高灵活性、内置的变量选择以及其非线性和非参数性质，已成为分析高维生存数据的重要方法。RSF 模型不需要满足特定的前提性假设，对于原始数据无特定要求，这受益于其非参数结构[14]。本文通过比较 Lasso-Cox 与 RSF 在生存数据上的分析性能，为今后生存数据分析的模型选择提供了参考依据，同时为临床上结直肠癌高危人群的筛选及 CRC 患者的治疗干预提供了理论依据。

本研究通过构建 Lasso-Cox 回归模型及 RSF 模

型共筛选出 13 个变量。C8G 是一种补体成分,其在外泌体中的成分可鉴别直肠癌患者对新辅助放疗的不同反应[15]。*HAND*1 作为一种可以异常甲基化的基因,可能在调节抑制生长或维持分化状态的其他基因的转录中发挥重要作用,Jin[16]等人认为结直肠癌中的 *HAND*1 发生甲基化。*HEPACAM*2 编码与免疫球蛋白超家族相关的蛋白质,该蛋白质在有丝分裂中起作用[17]。研究表明 *HEPACAM* 基因在结直肠癌肿瘤转移中起重要作用,且与结直肠癌的预后呈显著相关[18],这与本研究结果相同。在一项揭示结直肠癌患者肠道粘膜微生物组、代谢组和宿主 DNA 甲基化相关基因表达之间关联的多组学分析中,发现肿瘤组织中肠球菌属水平的降低与 *IGSF*9 的下调表达有关[19]。先前的研究报告了 *PCOLCE*2 在结直肠癌患者中的突变,但其具体机制尚不清楚,最近的研究发现,*PCOLCE*2 是结直肠癌的独立预后因素[20],这与我们的发现一致。更早的研究表明,*SCGB*2*A*1 是与化学耐药性和放射耐药性相关的结直肠癌的新型预后标志物[21]。此外 *SCGB*2*A*1 作为一种小的分泌蛋白的基因在子宫内膜癌和乳腺癌[22]等多种癌症中高表达,在胃癌中低表达[23]。Wu[24]等人从 IV 期 CRC 的原发性癌症组织中获得了测序和 RNA-seq 数据,并结合临床病理学和预后信息,运用多变量模型确定 *SFTA*2 在内的多个基因的 RNA 水平是患者预后的独立危险因素。此外,有研究证实 *TRPA*1 等通道的激活诱导结肠癌细胞中凋亡和氧化作用的增加,说明 *TRPA*1 等激活物可作为治疗结肠肿瘤的有效药物[25]。*CDK*5*R*2 在多种癌症的发展和诊断等方面存在重要作用[26]。但其与 *IGHV* 两种基因在 CRC 预后方面的研究较少,本研究结果对二者在 CRC 的预后作用上有一定提示作用。

研究采用了 VIMP 法分析 RSF 模型中变量重要性,有文献表明,虽然大多数情况下 VIMP 法和最小深度法所筛选出的变量类似,仍建议结合两种方法来进行变量的选择[27]。其次,通过计算一致性错误率进行模型准确度的比较,发现 Lasso-Cox 模型的预测错误率较 RSF 模型更低,准确度较好,这与陈哲[27]等人的研究结论相同。在模型校准度方面,RSF 模型在生存分析中的 IBS 小于 Lasso-Cox 模型,具有良好的模型校准度。因此,当数据满足 PH 假定且数据维度远大于样本量时,Lasso-Cox 分析更具优势;当数据不满足传统生存分析的假设前提时,RSF 模型作为一种模型校准度较好的生存分析方法可用于基因数据的预后分析。再次,本研究采用线性模型及经验贝叶斯方法的方法筛选差异表达基因,并将筛选出的基因用于 RSF 模型的构建,能够根据共同调节的基因组或高阶表达特征分析表达谱,提高了基因表达差异生物学解释的可靠性,使得 RSF 模型分析出的预后影响基因更加具有说服力。本研究仍存在不足,本研究基于在线基因数据库的回顾性分析,只进行了内部验证。

综上所述,RSF 模型在结直肠癌患者基因数据的预后分析中是可行的,且在模型校准度方面优于 Lasso-Cox 模型。由于随机生存森林模型不受 PH 假定、对数线性假定等条件的约束[5],可以用于不满足 PH 假定,变量之间可能存在相关性的数据进行预后影响因素分析,为结直肠癌的预后分析提供科学的依据。

参 考 文 献

[1] 徐萌,蔡美娟.CTC、MDSCs 及 S100A9 水平对Ⅱ~Ⅲ期结直肠癌术后临床预后的评估价值[J].中国现代普通外科进展,2022,25(7):529-533.

[2] Morvan L, Carlier T, Jamet B, et al. Leveraging RSF and PET images for prognosis of multiple myeloma at diagnosis[J]. Int J Comput Assist Radiol Surg,2020,15(1):129-139.

[3] Weng W, Zhang Z, Huang W, et al. Identification of a competing endogenous RNA network associated with prognosis of pancreatic adenocarcinoma[J]. Cancer Cell Int,2020,20:231.

[4] Kay R. Goodness of fit methods for the proportional hazards regression model: a review[J]. Rev Epidemiol Sante Publique,1984,32(3-4):185-198.

[5] Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests[J]. Annals of Applied Statistics,2008,2(3):841-860.

[6] 邓建新,单路宝,贺德强,等.缺失数据的处理方法及其发展趋势[J].统计与决策,2019,35(23):28-34.

[7] 杨耀,李四海.基于对称不确定性和 Lasso 的基因数据特征选择算法[J].信息技术与信息化,2022,262(1):8-11.

[8] Yosefian I, Farkhani EM, Baneshi MR. Application of Random Forest Survival Models to Increase Generalizability of Decision Trees: A Case Study in Acute Myocardial Infarction[J]. Comput Math Methods Med,2015,2015:576413.

[9] Ishwaran H. Variable importance in binary regression trees and forests[J]. Electron J Stat,2007,1:519-537.

[10] 李森,罗天娥,郭强,等.随机生存森林模型在肺癌患者预后分析中的应用[J].中国卫生统计,2021,38(3):327-331.

[11] Wang X, Gong G, Li N, et al. Detection Analysis of Epileptic EEG Using a Novel Random Forest Model Combined With Grid Search Optimization[J]. Front Hum Neurosci,2019,13:52.

[12] Adham D, Abbasgholizadeh N, Abazari M. Prognostic Factors for Survival in Patients with Gastric Cancer using a Random Survival Forest[J]. Asian Pac J Cancer Prev,2017,18(1):129-134.

[13] Kang L, Chen W, Petrick NA, et al. Comparing two correlated C indices with right-censored survival outcome: a onE-shot nonparametric approach[J]. Stat Med,2015,34(4):685-703.

[14] Mohammed M, Mboya IB, Mwambi H, et al. Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data[J]. PLoS One, 2021, 16(12):e0261625.

[15] Strybel U, Marczak L, Zeman M, et al. Molecular Composition of Serum Exosomes Could Discriminate Rectal Cancer Patients with Different Responses to Neoadjuvant Radiotherapy[J]. Cancers (Basel),2022,14(4):993.

[16] Jin B, Yao B, Li JL, et al. DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer[J]. Cancer Res,2009,69(18):7412-7421.
[17] Moh MC, Zhang T, Lee LH, et al. Expression of hepaCAM is downregulated in cancers and induces senescencE-like growth arrest via a p53/p21-dependent pathway in human breast cancer cells[J]. Carcinogenesis,2008,29(12):2298-2305.
[18] Huang Z, Yang Q, Huang Z. Identification of Critical Genes and Five Prognostic Biomarkers Associated with Colorectal Cancer[J]. Med Sci Monit,2018,24:4625-4633.
[19] Wang Q, Ye J, Fang D, et al. Multi-omic profiling reveals associations between the gut mucosal microbiome, the metabolome, and host DNA methylation associated gene expression in patients with colorectal cancer[J]. BMC Microbiol,2020,20(Suppl 1):83.
[20] Yao H, Li C, Tan X. An age stratified analysis of the biomarkers in patients with colorectal cancer[J]. Sci Rep,2021,11(1):22464.
[21] Munakata K, Uemura M, Takemasa I, et al. SCGB2A1 is a novel prognostic marker for colorectal cancer associated with chemoresistance and radioresistance[J]. Int J Oncol,2014,44(5):1521-1528.
[22] Zhang L, Yan X, Yu S, et al. LINC00365-SCGB2A1 axis inhibits the viability of breast cancer through targeting NF-κB signaling[J]. Oncol Lett,2020,19(1):753-762.
[23] Yan XY, Zhang JJ, Zhong XR, et al. The LINC00365/SCGB2A1 (Mammaglobin B) Axis Down-Regulates NF-κB Signaling and Is Associated with the Progression of Gastric Cancer[J]. Cancer Manag Res,2020,12:621-631.
[24] Wu B, Yang J, Qin Z, et al. Prognosis prediction of stage IV colorectal cancer patients by mRNA transcriptional profile[J]. Cancer Med, 2022,11(24):4900-4912.
[25] Kaya MM, Kaya İ, Nazıroğlu M. Transient receptor potential channel stimulation induced oxidative stress and apoptosis in the colon of mice with colitis-associated colon cancer: modulator role of Sambucus ebulus L[J]. Mol Biol Rep,2023,50(3):2207-2220.
[26] Pérez-Morales J, Mejías-Morales D, Rivera-Rivera S, et al. Hyperphosphorylation of Rb S249 together with CDK5R2/p39 overexpression are associated with impaired cell adhesion and epithelial-to-mesenchymal transition: Implications as a potential lung cancer grading and staging biomarker[J]. PLoS One,2018,13(11):e0207483.
[27] 陈哲, 许恒敏,李哲轩,等.随机生存森林:基于机器学习算法的生存分析模型[J].中华预防医学杂志,2021,55(1):104-109.

(责任编辑:郭海强)

(上接第 531 页)

[24] 张云权, 朱耀辉, 李存禄, 等. 广义相加模型在 R 软件中的实现[J]. 中国卫生统计, 2015, 32(6): 1073-1075.
[25] Hou W, Li Z, Zhang Y, et al. Using support vector regression to predict PM10 and PM2.5[J]. IOP Conference Series: Earth and Environmental Science, 2014: 012268.
[26] Efron B, Tibshirani RJ. An Introduction to the Bootstrap[J]. Chapman and Hall/CRC,1994.
[27] 饶克勤. 卫生统计方法与应用进展[M]. 北京: 人民卫生出版社, 2008.
[28] Politis DN, Romano JP. The Stationary Bootstrap[J]. Journal of the American Statistical Association, 1994, 89(428): 1303-1313.
[29] Martin MA. An Introduction to Bootstrap Methods with Applications to R[J]. Australian & New Zealand Journal of Statistics, 2012(2): 54.
[30] Flanders WD, Klein M, Darrow LA, et al. A Method for Detection of Residual Confounding in Time-series and Other Observational Studies[J]. Epidemiology, 2011, 22(1): 59-67.
[31] 阚慧. 阴性对照法:原理、方法及应用[J]. 中华流行病学杂志, 2020, 41(4): 5.
[32] Arnold BF, Ercumen A. Negative Control Outcomes: A Tool to Detect Bias in Randomized Trials[J]. JAMA, 2016, 316(24): 2597-2598.
[33] Basu R, Feng WY, Ostro BD. Characterizing temperature and mortality in nine California counties[J]. Epidemiology, 2008, 19(1): 138-145.
[34] Zanobetti A, Schwartz J. Temperature and mortality in nine US cities [J]. Epidemiology, 2008, 19(4): 563-570.
[35] Chen R, Yin P, Meng X, et al. Fine Particulate Air Pollution and Daily Mortality: A Nationwide Analysis in 272 Chinese Cities[J]. Am J Respir Crit Care Med, 2017, 196(1): 73-81.
[36] 夏钟, 王欣童, 郁莎燕, 等. 应用 Meta 分析研究中国不同区域 PM2.5 污染与人群非意外总死亡率的关系[J]. 环境污染与防治, 2019(8):5.
[37] Bae S, Lim YH, Hong YC. Causal association between ambient ozone concentration and mortality in Seoul, Korea[J]. Environmental Research, 2020, 182(3): 109098.
[38] Glymour MM, Tchetgen E, Robins JM. Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions[J]. American Journal of Epidemiology, 2012, 176(4): 456.

(责任编辑:郭海强)