Statistics and Application 统计学与应用, 2022, 11(4), 892-908
Published Online August 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.114093

Hans汉斯

基于数据挖掘的肝内胆管癌预后分析

韩元全，雷　杰，陈　浪

重庆理工大学理学院，重庆

收稿日期：2022年7月25日；录用日期：2022年8月5日；发布日期：2022年8月17日

摘　要

本文将采用美国癌症数据库SEER的患者数据，基于三种预测模型对肝内胆管癌(ICC)患者临床病理数据进行分析，完成对生存时间以及生存状态的预测。首先选取了年龄、性别、种族、T分期、N分期、M分期等变量完成了生存曲线的刻画，通过单因素分析和多因素分析确定了Cox风险回归模型的变量及系数，构建了Cox预测模型。接着完成了机器学习算法的建模，首先是梯度提升法建立的模型，选取了一共12个变量并筛选出了5个最重要的独立变量完成预测。然后是BP神经网络模型的应用，排除了种族等不重要的分类变量，加入了对ICC预后影响更明显的变量进入模型。通过对三个模型拟合效果对比得出相应结论。最后针对上述研究结果提出合理的建议。

关键词

肝内胆管癌，Cox比例风险回归模型，梯度提升，神经网络，美国癌症数据库SEER

Prognostic Analysis of Intrahepatic Cholangiocarcinoma Based on Data Mining

Yuanquan Han, Jie Lei, Lang Chen

Faculty of Science, Chongqing University of Technology, Chongqing

Received: Jul. 25th, 2022; accepted: Aug. 5th, 2022; published: Aug. 17th, 2022

Abstract

This article will use the patient data of the American cancer database SEER to analyze the clinicopathological data of patients with intrahepatic cholangiocarcinoma (ICC) based on three prediction models to complete the prediction of survival time and survival status. First, variables such as age, gender, race, T stage, N stage, and M stage were selected to describe the survival curve. The variables and coefficients of the Cox hazards regression model were determined through univa-

文章引用: 韩元全, 雷杰, 陈浪. 基于数据挖掘的肝内胆管癌预后分析[J]. 统计学与应用, 2022, 11(4): 892-908.
DOI: 10.12677/sa.2022.114093

riate analysis and multivariate analysis, and the Cox prediction model was constructed. Then, the modeling of the machine learning algorithm is completed. The first is the model established by the gradient boosting method. A total of 12 variables are selected and the 5 most important independent variables are selected to complete the prediction. Then there is the application of BP neural network model, which excludes unimportant categorical variables such as race, and adds variables that have a more obvious impact on the prognosis of ICC into the model. Corresponding conclusions are drawn by comparing the fitting effects of the three models. Finally, reasonable suggestions are put forward based on the above research results.

Keywords

Intrahepatic Cholangiocarcinoma, Cox Proportional Hazards Regression Model, Gradient Boosting, Neural Network, US Cancer Database SEER

1. 引言

在全球范围内，肝癌以往一直是让医学界非常棘手的癌症，根据世界卫生组织国际癌症研究机构(IARC)发布的 2020 年全球癌症发病情况分析报告中指出肝癌的发病率在全球第六位，而其死亡率则高达第三。尽管在科技日益发达的今天，我们也丝毫不能小觑肝癌对人类健康的影响。目前肝癌新发病例的发生率为每年每 100,000 名男女 9.5 例。死亡率为每年每 100,000 名男女 6.6 人。

肝癌是指原发于肝脏组织细胞的恶性肿瘤。由于肝脏里有肝细胞、胆管细胞、间质细胞等等，所以其实肝癌可以细分为好几种[1]。主要有两种：一种是来源于肝细胞的肝细胞癌，占到 80%以上；另一个就是来源于胆管细胞的肝内胆管细胞癌(因此本质上是胆道肿瘤)，占到 10%~15%。近 30 年肝内胆管癌(ICC)的发病率在全世界范围内呈明显上升趋势。肝内胆管癌发病隐匿，极易侵犯肝脏周围器官、组织和神经，发生淋巴结和肝外远处转移，大部分病人确诊时通常已处于晚期，缺乏有效治疗方法。对于部分早期肝内胆管癌病人，肝切除治疗已获得广泛肯定。然而，即使行根治性切除术，肝内胆管癌术后仍然极易复发和转移，病人术后 5 年总体生存率为 25%~40%，预后远差于肝细胞癌[2]。

通常来说，诊断时的癌症阶段也就是体内癌症的程度，决定了治疗方案，并对生存期的长短有很大的影响。一般来说，如果癌症只在它开始的身体部位被发现，它就会被定位(有时称为第一阶段)。如果它已经扩散到身体的不同部位，则该阶段是区域性的或远处的。肝内胆管癌越早被发现，一个人在被诊断出 5 年后存活的机会就越大。对于肝内胆管癌，43.9%在局部分期时被诊断出来。而局限性肝内胆管癌的年相对生存率为 36.1%。这也是相比其他癌症阶段最高的生存率了。

随着医学技术的进步，肝癌现有的治疗方法日新月异。目前肝内胆管癌的治疗方法主要包括：手术治疗、肝移植、消融治疗、放射治疗、化学疗法、靶向疗法、免疫疗法[3]。根据病人的病情选择不同的治疗方法能最大化有益于患者的身体健康，所以如何选择最合适的疗法显得尤为关键。

当得知了患者所处的癌症阶段后，通过病情以及患者信息判断得出最有效的治疗方法显得尤为重要。鉴于上述内容，本文选取了 2538 例来自于 SEER 数据库中的肝内胆管癌患者信息进行生存分析研究，通过建立预测模型基于患者不同的病情等信息对患者生存时间和生存状态进行预测。

2. 数据与方法

2.1. 数据来源

本文的数据是来自于 SEER 即美国国家癌症研究所监测，流行病学和最终结果数据库(https://seer.cancer.gov)，这个数据库主要包括了人口统计数据、患者的个人临床信息、肿瘤种类、治疗方式以及生存状态等。本文通过 SEER*Stat 软件来获取相关所需数据模块，在案例表格板块(Case Listing Session)下载 2005 年至 2014 十年间诊断的所有肝内胆管癌病例，共计 2538 例患者信息。其中，患者纳入选择的标准如下：1) 患者确诊年份在 2005 年至 2014 年之间；2) 根据国际疾病分类肿瘤学专辑第三版(ICD-O-3)原发部位选择肝内胆管；3) 镜下证实的原发性肝内胆管癌；4) AJCC 的 T 分期从 T1 到 TX；5) 年龄大于 20 岁。最终本文数据纳入的变量包括：性别、种族、年龄、美国癌症联合委员会(AJCC) T、N、M 分期、肿瘤大小、区域淋巴结数量、切除的淋巴结总数、淋巴结转移数目、总体转移数目、肿瘤总数、生存时间、生存状态。

2.2. 数据说明与预处理

本文一共选取了 14 个变量，其中生存时间和生存状态作为本次模型预测对象，另外 12 个变量均为自变量，上述变量选择都基于常用的临床病理参数。由于下载的年龄数据被细分为 5 年一个阶段，经过初步数据分析，大部分患者年龄在 50 岁以上，其中 70 岁以上的患者也占多数，于是将年龄分为 3 个阶段：20~49 岁，50~69 岁，70 岁以上。具体变量说明如表 1 所示。

Table 1. Variable description and assignment representation
表 1. 变量说明及赋值表示

变量	变量名	英文名	变量类型及赋值
X1	年龄	Age	20~49 years；50~69 years；70+ years
X2	性别	Sex	Male；Female
X3	种族	Race	White；Black；Other
X4	T 分期	AJCC.T	T0；T1；T2；T3；T4；T5；TX
X5	N 分期	AJCC.N	N0；N1；NX
X6	M 分期	AJCC.M	M0；M1；MX
X7	区域淋巴结数量	Regional nodes positive	数值
X8	切除的淋巴结总数	Regional nodes examined	数值
X9	淋巴结转移数目	CS lymph nodes	数值
X10	总体转移数目	CS mets at dx	数值
X11	肿瘤总数	Total number of patient	数值
X12	肿瘤大小	CS tumor size	数值

在整理数据的过程中存在一些缺失值或者空值，考虑到数据较多的变量，则用其他数据的平均值或者 0 代替填入，对于丢失数据较少的变量，则直接删除掉该病人的所有信息。另外，对于生存结局变量，令死亡 ＝2，存活 ＝1。为了测试模型精度，将数据集 70%划分为训练集，30%为测试集。

2.3. 模型选择

Cox 模型全称比例风险回归模型，以生存结局和生存时间为因变量，它可同时分析多个因素对生存时间的影响，分析存在部分信息缺失的资料且不要求资料的分布类型。因为我们将进行的是对肝内胆管癌预后的多因素分析，所以毫无疑问 Cox 模型是一个不错的选择。

事实上，目前为止 Cox 模型已经广泛应用于生存分析中，尽管有不错的效果，但随着计算机辅助智能诊疗技术的飞速发展，机器学习方法可以通过从大量、丰富的医疗数据中学习，不依赖固定的、简易的建模规则，并在多种临床情境下展现出巨大优势。但是由于多数机器学习算法的内部结构并不透明，在可解释性方面远差于传统的回归模型即这里谈到的 Cox 模型，因此机器学习方法并不易于被临床医师理解与接受，更不用说应用了。梯度提升机(GBM)是机器学习分析中常用的方法，该算法由大量简单的决策树集合而成，具有较高的可解释性。因此，本文将在使用 Cox 模型的基础上，完成对梯度提升机模型的构建，旨在对比模型效果，为优化肝内胆管癌预后评估提供新思路与新方法。

同时，我们提到由于机器学习方法的可解释性受到质疑，所以本文在应用可解释性较高的梯度提升机的同时想到了建立另一可解释性不高的机器学习模型加以对比，即 B-P 神经网络模型。

3. 模型预测结果实证分析

本章主要是对选取的癌症病人数据进行生存分析，在利用 Cox 比例风险回归模型做预测的基础上，通过对机器学习方法的探索，采用梯度提升机以及神经网络模型对病人的生存时间和生存状态进行预测，同时分析了不同因素对其生存概率的影响。

3.1. Cox 比例风险回归模型

3.1.1. K-M 生存曲线

1) 不同性别的生存特征

不同性别的患者的生存概率是具有差异性的($p < 0.05$)，如图 1。可以看出其中女性的生存概率总体上是高于男性的，这也是一个可以值得研究的点。在生存时间少于 5 个月的患者中生存概率一样即性别差异是不明显的，同样当生存时间超过 120 个月时性别也不具有差异了。男性的生存时间中位数为 5 个月，而女性生存时间中位数为 6 个月，这也说明男性平均生存时间要短于女性。

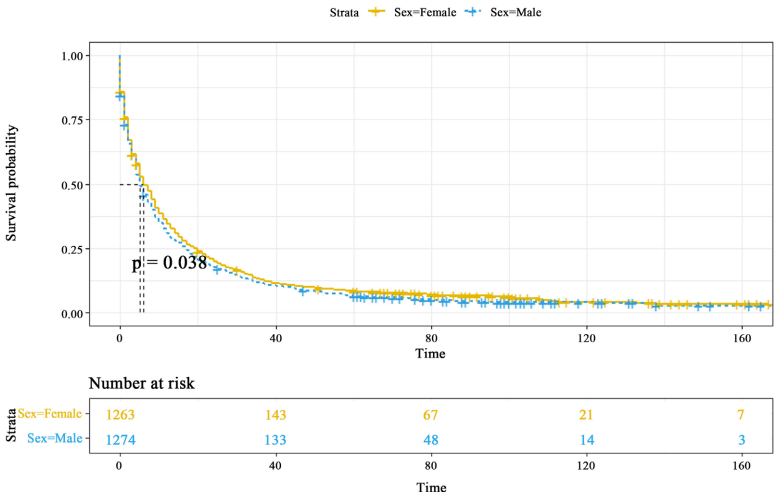

Figure 1. K-M survival curves of different genders
图 1. 不同性别的 K-M 生存曲线

2) 同种族的生存特征

人种方面主要分为白种人和黑种人以及其他种族。由图 2 的生存曲线可以看出白种人的生存概率是明显高于黑种人和其他种族的，尤其是在生存时间在 0 至 60 个月区间的时候。但同时也应关注到另一事实就是，虽然白种人生存概率高，但是患病人数远远高于黑种人和其他种族，在抽取的样本中，白种人患病数是其他种族的 4 倍，这也与美国本地种族比例有密切关系。另外，白种人生存概率高的同时意味着黑种人的死亡风险比率过高。

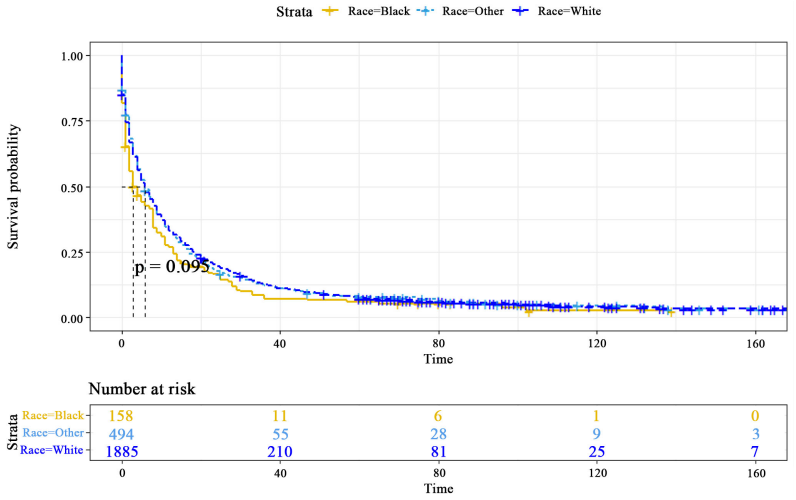

Figure 2. K-M survival curves of different races
图 2. 不同种族的 K-M 生存曲线

3) 同年龄的生存特征

对于不同年龄段的患者，如图 3，本文根据样本数据量大小将年龄分为 3 个段：20~49 岁，50~69 岁，70 岁以上。经过年龄段划分，明显看出 3 个年龄段的生存概率截然不同，首先年龄从小到大，其每个时期的生存概率均呈现从大到小，其次生存时间中位数方面，老年人的生存时间明显大打折扣，仅为 3 个月左右，远低于年轻人的预后表现。再者老年人的生存概率在 60 个月后基本为 0。

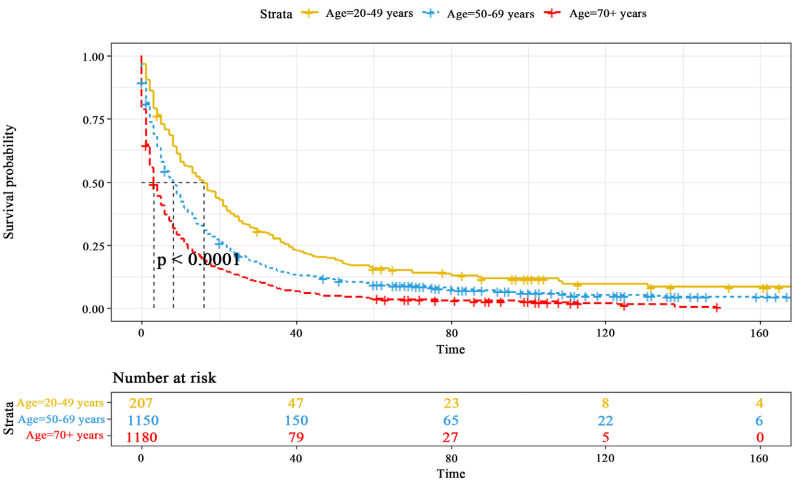

Figure 3. K-M survival curves of different ages
图 3. 不同年龄的 K-M 生存曲线

4) 同 T 分期的生存特征

TNM 分期系统是目前国际上最为通用的肿瘤分期系统，也是临床上进行恶性肿瘤分期的标准方法。图 4 中由于 T0 分期的患者数过少不予考虑，其余生存率最高的是 T1 分期的患者，其次是 T2 分期、T3 分期、T4 分期生存率从高到低，这与其肿瘤分期大小也息息相关。另外，TX 分期的患者肿瘤情况未知，但生存率是最低的，某种程度上说明其肿瘤程度相对严重以至于死亡率高于其他 T 分期患者。

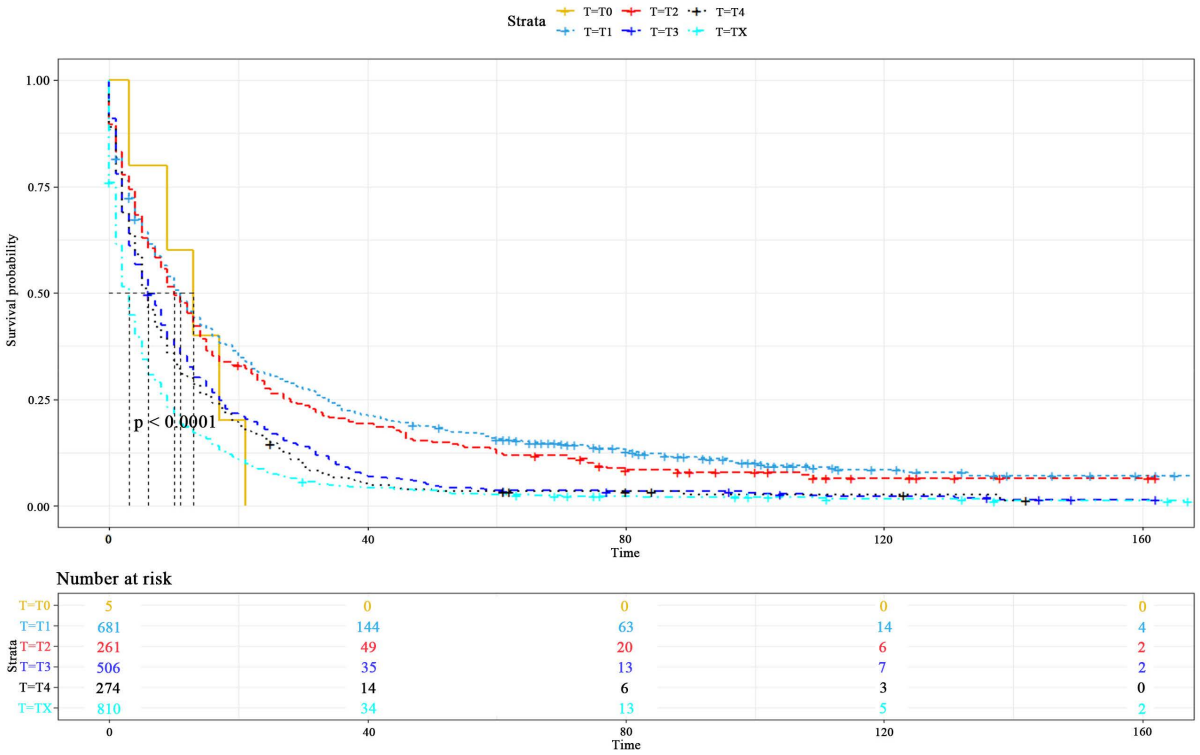

Figure 4. K-M survival curves of different T stages
图 4. 不同 T 分期的 K-M 生存曲线

5) 同 N 分期的生存特征

本次选取的数据量中 N 分期的患者只有 3 种类型，如图 5，其中 N0 表示没有区域淋巴结转移的患者，所以明显看到处于 N0 分期的患者生存率是最高的，而 N1 分期的患者生存率较 N0 分期的患者大幅度下降，在 40 个月后生存概率基本为 0。NX 分期表示区域淋巴结未知状况，同 T 分期一样，生存率远低于另外两分期状态。也说明未知状况下的淋巴结转移状态是更差的。

6) 同 M 分期的生存特征

M 分期表示的是远端转移的情况，图 6 中 M0 分期表示无远处转移，显而易见，其生存概率明显高于处于其他分期的患者，M1 分期的患者表示有远端转移的迹象，生存概率也大打折扣，与位于 MX 分期即未知远端转移情况的患者生存概率不相上下，在 40 个月后其死亡率也基本接近 100%。

3.1.2. 单因素与多因素分析

本节对数据进行单因素和多因素分析，分析的变量与分析生存曲线的变量相同，由于这里主要是对分类变量进行分析，故并未将区域淋巴结数量、切除的淋巴结总数等数值变量纳入 Cox 风险模型的构建中。此处的因变量则是生存时间和生存状态。

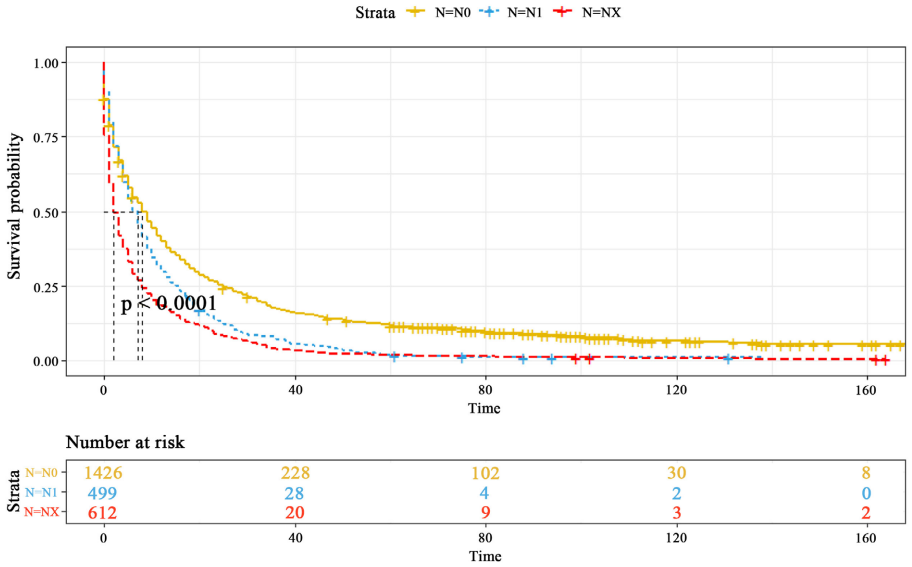

Figure 5. K-M survival curves of different N stages
图 5. 不同 N 分期的 K-M 生存曲线

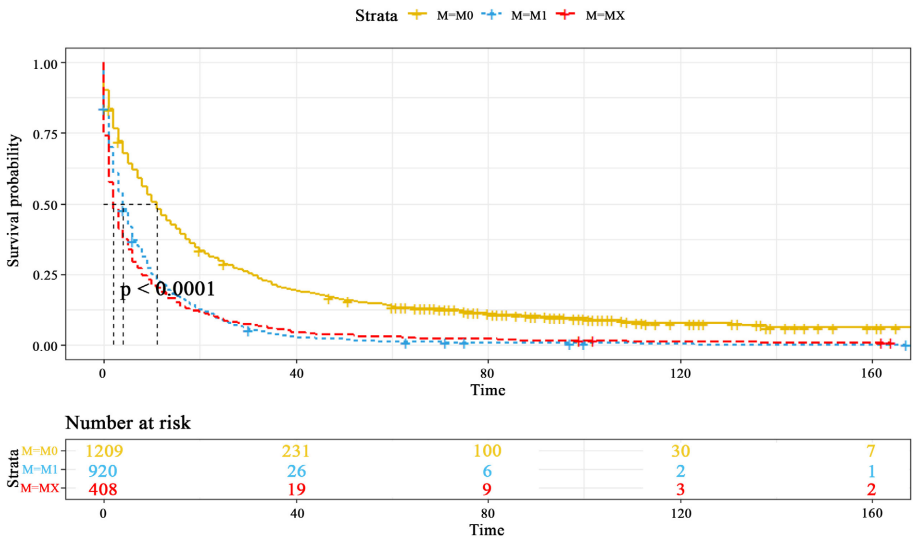

Figure 6. K-M survival curves of different M stages
图 6. 不同 M 分期的 K-M 生存曲线

从单因素分析表 2 中可以看出，年龄、性别、种族、N 分期、M 分期以及肿瘤总数 6 个变量均通过了显著性检验，只有 T 分期中所有变量均为通过显著性检验($p > 0.05$)。首先观察年龄变量，得出年龄越大，其死亡风险越高，并且其变化率是迅速增长的，大于 70 岁的老人的死亡风险比 20~49 岁的年轻人高了一倍不止(HR = 2.099)。性别方面男性确实比女性死亡风险大，但是差异不是特别明显。对于种族来说，白种人总体死亡风险明显小于黑种人以及其他种族，这也说明白种人的预后是更好的。T 分期中表示肿瘤大小与死亡风险之间没有较强联系，这也许与肝内胆管癌的特殊性有关，在本文中不加以讨论。N 分期和 M 分期都是关于肿瘤转移的信息，其中处于 N1 分期的患者死亡风险是 N0 分期患者的 1.37 倍(HR = 1.37)，而处于 NX 分期的患者是 N0 患者的 1.95 倍(HR = 1.95)。对于 M 分期而言，处于 M1 和 MX 分期的患者的死亡风险较 M0 患者而言均高达 2 倍左右。

在多因素分析表中，与单因素分析中变量影响基本一致，除了 T 分期没有通过显著性检验之外，其他变量均通过了显著性检验。总体来看，年龄越大死亡风险越高，并且不同年龄段的死亡风险有显著差异，对比单因素，死亡风险均有提升。而同时，其他多个变量的死亡风险较单因素都有不同程度的下降。

Table 2. Univariate and multivariate analysis tables
表 2. 单因素与多因素分析表

	单因素分析		多因素分析	
	HR	p 值	HR	p 值
年龄				
20~49	参考		参考	
50~69	1.409	<0.01	1.4925	<0.01
>70	2.099	<0.01	2.3832	<0.01
性别				
女性	参考		参考	
男性	1.089	0.0376	1.1066	0.0139
种族				
黑种人	参考		参考	
白种人	0.8295	0.0285	0.6991	<0.01
其他	0.8284	0.0455	0.7347	<0.01
T 分期				
T0	参考		参考	
T1	0.7373	0.498	0.9330	0.88
T2	0.7986	0.619	1.0884	0.85
T3	1.1086	0.819	1.3309	0.53
T4	1.1779	0.717	1.3244	0.53
TX	1.6258	0.279	1.5082	0.36
N 分期				
N0	参考		参考	
N1	1.3785	<0.01	1.1806	<0.01
NX	1.9537	<0.01	1.2614	<0.01
M 分期				
MO	参考		参考	
M1	1.9695	<0.01	1.6413	<0.01
MX	2.1809	<0.01	1.4156	<0.01
肿瘤总数	0.8413	<0.01	0.8071	<0.01

3.1.3. 预测模型的构建

根据多因素分析可以完成对 Cox 预测模型的构建，将未通过显著性检验的变量即 T 分期变量移除掉 ($p > 0.05$)，纳入变量为：性别、年龄、种族、N 分期、M 分期以及原发或恶性肿瘤总数。确定的模型如下：

$$h\left(t\middle|X\right)=h_0\left(t\right)\exp\left\{\begin{matrix}-0.21X_1-0.07X_{21}+0.08X_{22}+0.28X_{23}+0.28X_{24}+0.41X_{25}\\+0.16X_{31}+0.23X_{32}+0.49X_{41}+0.34X_{42}-0.30X_{51}-0.35X_{52}\\+0.40X_{61}+0.86X_{62}+0.10X_7\end{matrix}\right\}$$

其中 X_1 表示原发或恶性肿瘤总数， X_2 表示 T 分期的五个阶段， X_3 表示 N 分期的两个阶段， X_4 表示 M 分期的两个阶段， X_5 表示种族， X_6 表示年龄， X_7 表示性别中的男性。

基于上述模型，对其进行显著性检验，模型的显著性检验结果如下表 3。可以看到三种检验方式的 p 值都远低于 0.05，这说明模型通过了显著性检验，而且模型整体效果不错，表示这 6 个变量能够相对合理的对因变量即生存时间或生存状态完成预测。

Table 3. Significance test table
表 3. 显著性检验表

检验方式	卡方值	自由度	p 值
Likelihood ratio test	587.5	10	p ≤ 2e−16
Wald test	575.1	10	p ≤ 2e−16
Score (logrank) test	594.6	10	p ≤ 2e−16

接下来基于上述模型我们可以完成列线图的绘制。由于患者生存时间有限，我们希望知道患者能存活到某个时间点之后的概率，于是本文完成了对患者诊断后生存 6 个月、18 个月和 36 个月的生存概率的预测。这里生存截止时间点是基于本次论文的数据生存时间分布来选择的，其中有 50%以上的人生存时间不足 6 个月，70%以上的人生存时间不足 18 个月，90%的人生存时间不足 36 个月。

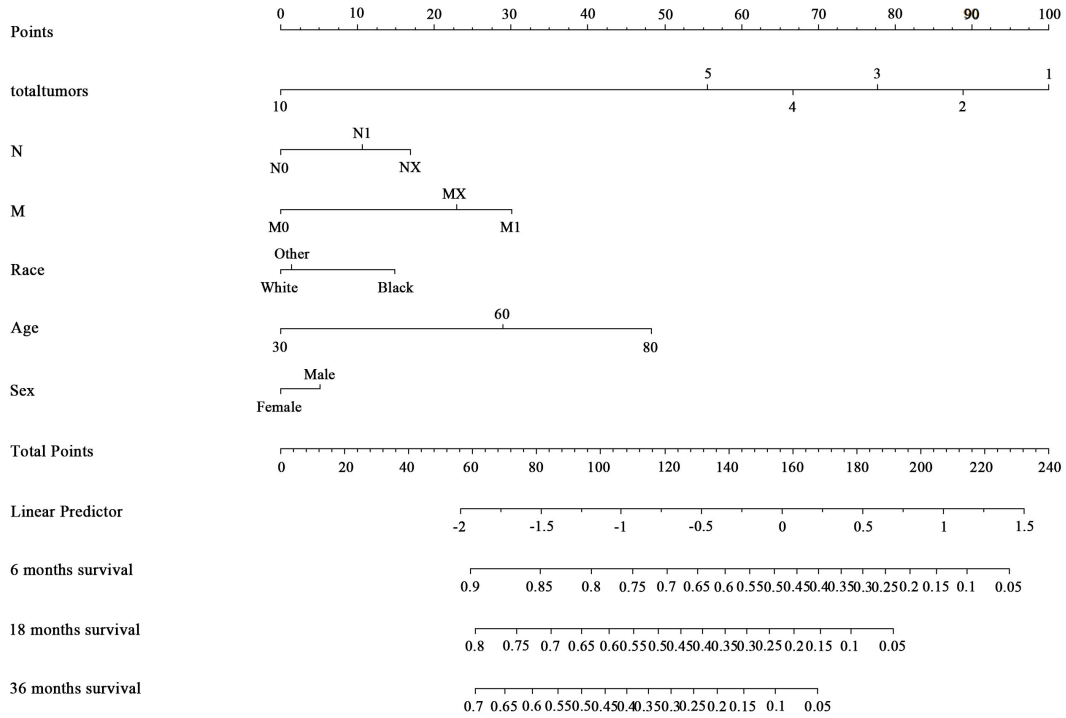

Figure 7. Patient survival prediction nomogram
图 7. 患者生存率预测列线图

图 7 给出了患者不同特征信息对应的不同生存时间的生存概率，上图表示每个患者不同的特征信息对应第一行的分数刻度，每一个预测变量都有对应的分数，然后将所有预测变量的分数相加，根据总分数可以得到对应不同生存时间的概率。举个例子：一个 60 岁的女性患者，她是白种人，处于 N1 分期和 M1 分期，一共有 5 个肿瘤。那么他的分数为：30 + 0 + 0 + 10 + 30 + 55 = 125 分，那么可以得到她 6 个月的生存概率为 0.65，18 个月的生存概率为 0.4，36 个月的生存概率为 0.25。

3.1.4. 预测模型的验证

1) ROC 曲线

图 8 中左右两图分别是此次预测模型关于训练集和预测集的 ROC 曲线图。训练集的 ROC 曲线图对应的 6 个月、18 个月、36 个月 AUC 为 0.447、0.458、0.458，而测试集对应的 6 个月、18 个月、36 个月 AUC 为 0.363、0.274、0.274，整体预测效果差于训练集。两者的值都偏低，训练集与测试集表明模型预测效果不佳，总体预测效果较差。

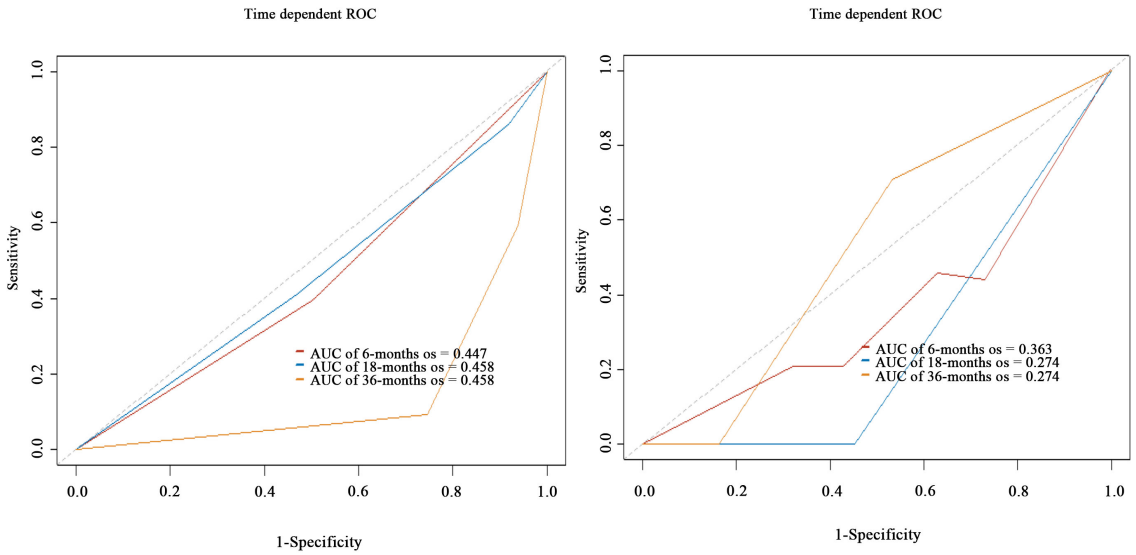

Figure 8. Cox training set and test set patient ROC curve
图 8. Cox 训练集与测试集患者 ROC 曲线图

2) 校准曲线

图 9~11 分别描述的是患者生存时间分别为 6 个月、18 个月、36 个月的校准曲线图。比较明显的是患者 6 个月的校准曲线图较大程度上偏离了对角线，预测效果不佳，虽然预测结果波动不大，但是预测生存概率总体偏低太多。对于患者 18 个月的校准曲线可以看到，预测生存概率相对贴合，但预测总体生存概率仍然偏低。相反的是患者 36 个月的校准曲线图，整体预测概率偏高。

综上可以看出，不论是短期预测还是长期预测效果都不会太好，尽管长期预测效果看上去还过得去，但总体也比较一般。考虑到本次数据选取量并不大，并且偶然性较大，后续模型将仅对患者生存 36 个月时间进行长期预测。

3.2. 梯度提升机模型

3.2.1. GBM 模型构建与验证

本节将基于训练集患者的数据对患者生存结局进行总体预测，共 12 个临床病理特征纳入 GBM 模型

分析。通过顺序后向搜索法，GBM 算法从中筛选出 5 个重要特征：年龄、原发或恶性肿瘤数目、肿瘤大小、区域淋巴结数量、切除的淋巴结总数(图 12)。

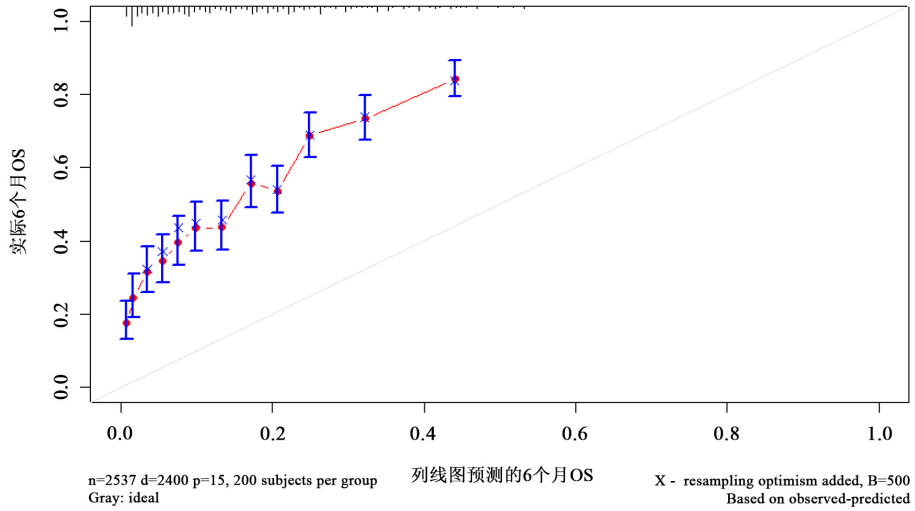

Figure 9. 6-month calibration curve graph for the patient
图 9. 患者 6 个月的校准曲线图

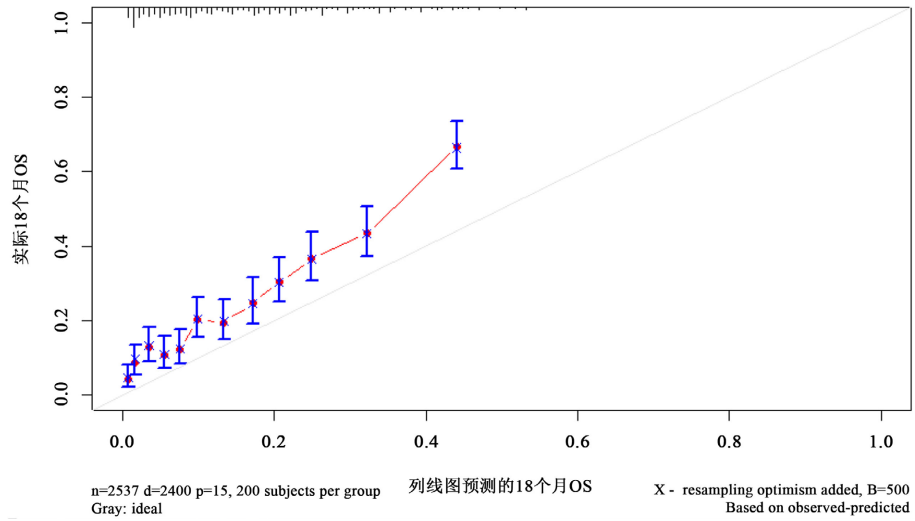

Figure 10. 18-month calibration curve graph for the patient
图 10. 患者 18 个月的校准曲线图

基于上述变量构成的模型计算后可以得到最佳迭代次数，由图 13 可以看出迭代到第 839 次时，模型表现不再有进一步的提升，因此最佳迭代次数为 839。

3.2.2. GBM 模型评价

图 14 左右两图分别表示患者生存 36 个月的训练集以及测试集模型预测 ROC 曲线图。训练集的模型拟合效果要优于测试集，但总体差距不大，AUC 均超过 0.7，表明模型预测效果较好，整体区分度不错，与 Cox 风险回归模型相比，尽管只预测了 36 个月的生存数据，效果也明显好于前者，梯度提升法是一个更优选择在模型拟合方面。

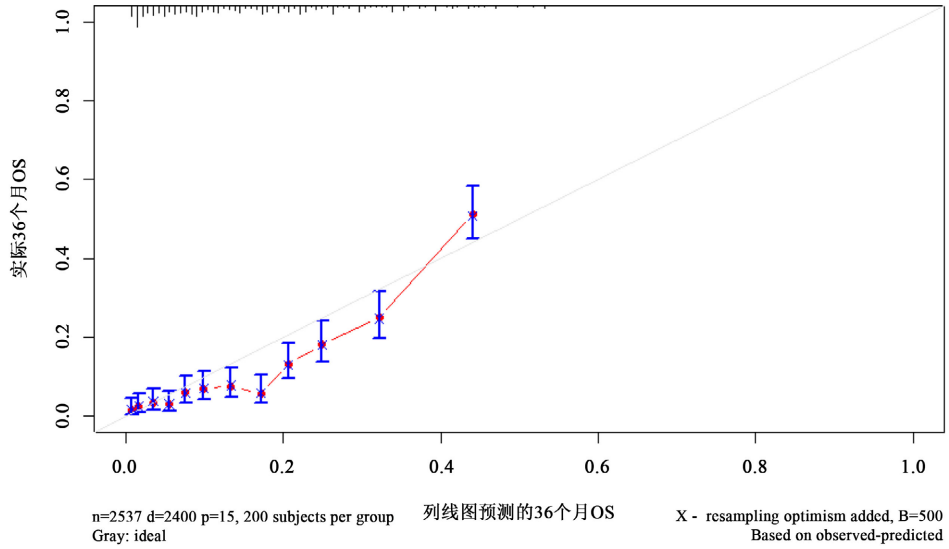

Figure 11. 36-month calibration curve graph for the patient
图 11. 患者 36 个月的校准曲线图

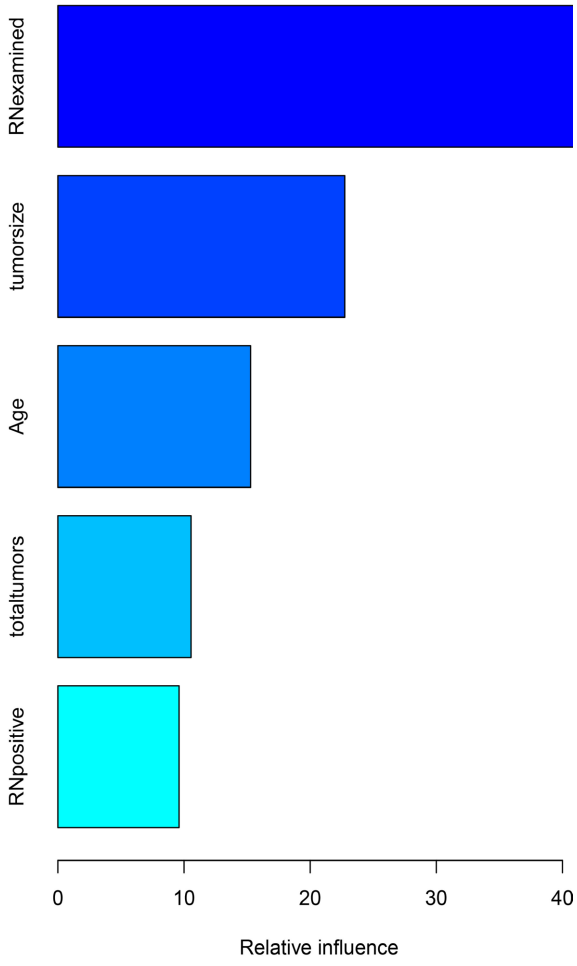

Figure 12. Variable relative importance
图 12. 变量相对重要性

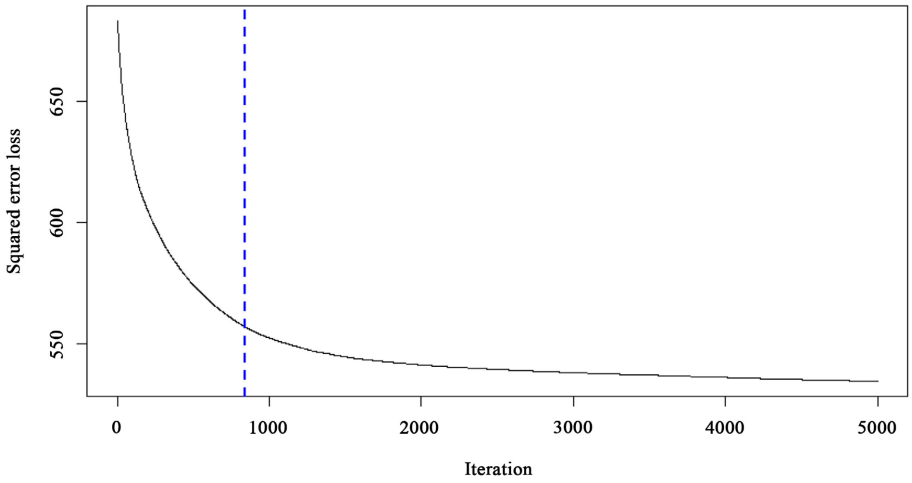

Figure 13. Iteration number error plot
图 13. 迭代次数误差图

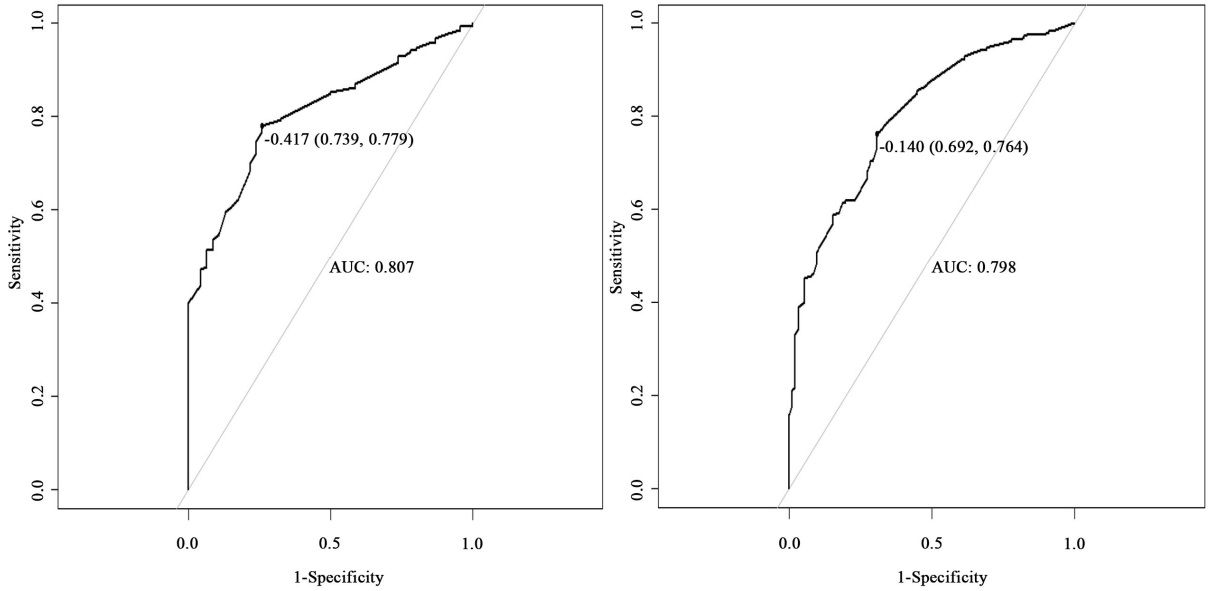

Figure 14. GBM training set and test set patient ROC curve
图 14. GBM 训练集与测试集患者 ROC 曲线图

3.3. 基于 B-P 算法的神经网络模型

与前两节的模型构建不同的是，神经网络模型的变量选择方面比较灵活，这里考虑的变量选择重点是在临床病理特征上，新纳入了肿瘤大小、区域淋巴结数量、切除的淋巴结总数、淋巴结转移数目、总体转移数目多个变量，在之前的多因素分析中，由于年龄也属于一个较强的影响因子，这里同样将其纳入预测变量中。因变量则是患者的生存状态：死亡或者存活。

3.3.1. 神经网络模型构建

图 15 中，可以清楚看到，本次模型一共有 7 个预测变量，这里的模型一共有两个隐藏层，其中一个隐藏层有 5 个节点，另一个有 3 个节点。图中的数字表示每个节点的权重，用以预测最终 36 个月的生存结局。

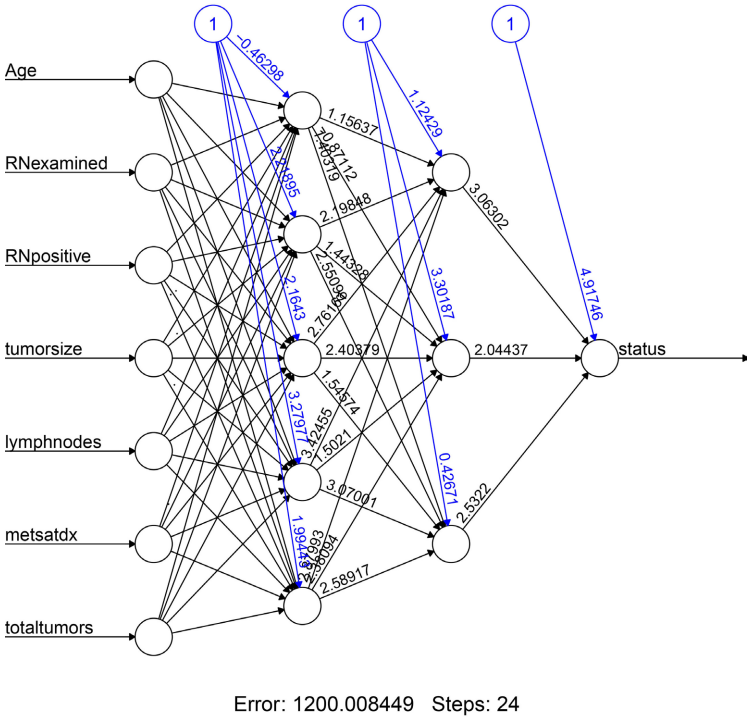

Figure 15. A neural network model for predicting patient survival at 36 months
图 15. 患者 36 个月的生存率预测神经网络模型

3.3.2. 神经网络模型评价

图 16 左右两图分别表示对患者生存 36 个月的训练集以及测试集模型预测 ROC 曲线图。在训练集方面，神经网络模型的预测表现很好，AUC 值为 0.841。测试集 AUC 的值为 0.816，略高于训练集的值，但是也足以说明模型的拟合效果不错，整体预测精度也较 Cox 风险回归模型和梯度提升模型大幅上升。

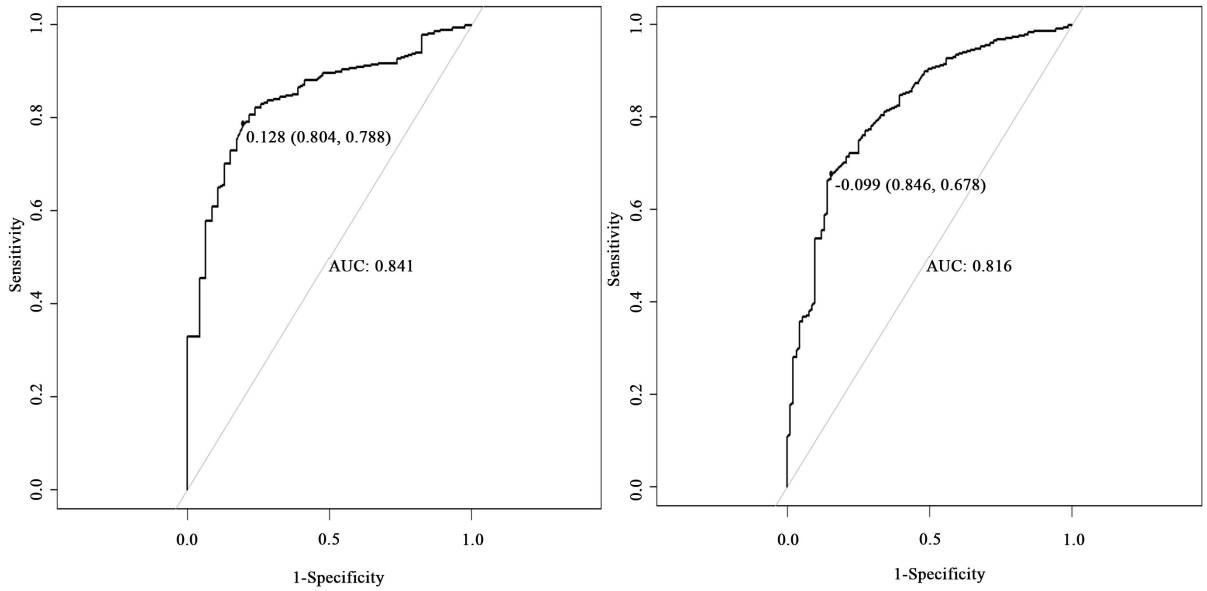

Figure 16. Neural network training set and test set patient ROC curve
图 16. 神经网络训练集与测试集患者 ROC 曲线图

3.4. 模型对比

本节一共运用了 3 个预测模型对患者的生存时间以及生存状态进行预测，Cox 风险回归模型中使用的大多数变量都是分类变量，例如性别、年龄、种族等，通过单因素与多因素分析以及 K-M 生存曲线分析每个变量不同类别的独立性差异，筛选出对模型影响最多的预测变量构建模型。最后得到的 Cox 风险回归模型 ROC 曲线以及对应的校准曲线表明，ROC 曲线下的 AUC 值实在偏低，远没有达到预期的预测目的，而校准曲线方面也说明离正确值偏离较多，模型整体还有待改进。

梯度提升法模型在变量选择的过程中将 12 个变量均纳入其中，模型筛选出了 5 个最重要的预测因子，并且给出了最佳迭代次数。由于考虑到进行短期预测效果并不会很好，所以只对 36 个月患者生存时间的生存概率进行了预测，跟预想的效果一样，对长期生存时间的预测是比较准的，相应 ROC 曲线的 AUC 达到了 0.7 以上，相较 Cox 风险回归模型而言，整体预测效果大幅提升，但是也考虑到这里的变量选择并不是十分令人满意，接着又利用神经网络模型完成了患者生存时间长期预测。

对于神经网络模型而言，变量的选择方面比较简单，由于其模型本身对变量不做过多限制，本文主要采用了临床上比较重要的几个病理特征作为变量，例如肿瘤大小，区域淋巴结数目等等，打包一起代入模型中进行预测。比较直观的效果就是对应的 ROC 曲线下 AUC 的值在训练集甚至达到了 0.8 以上，其拟合效果是毋庸置疑的，精度方面也是 3 个模型最高的。但是同时也应注意到神经网络模型在变量解释方面较 Cox 模型完整度上差了不少。

4. 总结与展望

4.1. 基于预测模型的讨论

精准预测预后对肝内胆管癌(ICC)病人的治疗决策具有重要意义。尽管诸多学者通过探索全新的基因或分子标记物用于改善预后评估和治疗选择，但因检测方法费时费力、价格昂贵且缺乏统一检测手段，因此该技术离临床广泛运用尚存一定距离。事实上，基于现有的临床数据构建一个简单的评分系统来实现精准的预后预测，仍是临床肿瘤学个体化治疗的首选参考工具。例如，临床医生已经开始使用简易的模型来评估胆囊癌病人接受辅助放化疗的获益程度。

本篇文章基于最常见的临床病理参数以及 SEER 数据库内 2538 例 ICC 病人的数据，构建并验证了全新的预后预测模型，结果显示：传统的 Cox 风险回归模型并不能满足现阶段部分预测需求，既往研究所构建的 ICC 预后预测模型多基于 Cox 回归建模策略。Cox 模型其假设协变量之间的相互作用是均匀的，不同协变量在风险函数中是独立的，但事实上 ICC 预后相关的因素之间存在复杂的交互关系。同时，Cox 回归分析必须在病人临床信息完整的情况下进行，因部分变量缺失而将病人除外后续分析会产生巨大的选择偏倚，这些都导致了本文章中 Cox 模型拟合效果不好。

此时，机器学习算法便大有可为。随着机器学习算法在临床研究中的不断深入，模型的可解释性以及对缺失数据的处理能力成为了机器学习研究中关注的重点。基于决策树的方法是机器学习算法中的一个大家族，可依据预测因子之间复杂的非线性关系来区分预后亚组。分类与回归树算法因其简洁和直观的模型结构，已被用于区分不同 ICC 病人的手术预后，但该算法的预测效能有限。梯度提升算法(GBM)是一种包含大量决策树的集成学习算法，即将若干个独立的分类与回归树整合成一个强分类器来实现精确和稳定的预测。因此 GBM 模型的内在结构是可拆解的，也便于临床医师理解 GBM 算法是如何实现高效预测的。此外，GBM 算法存在处理缺失值的内置功能，通过利用队列中的数据进行分类，而无须对缺失数据进行插补。这大大拓宽了可用于建模的数据集，减少了因删除存在缺失值的病人所导致的选择偏倚。本文的 GBM 模型拟合整体效果是很不错的，可以作为后续预测研究的一个较好选择。GBM 模型能

够准确预测 ICC 病人的预后。因此，本研究构建的 GBM 模型可为 ICC 病人临床决策提供重要参考。

BP 神经网络模型也作为非常经典有用的一个机器学习算法，其具有逼近效果好，计算速度快，不需要建立数学模型，精度高以及理论依据坚实，推导过程严谨，所得公式对称优美，具有强非线性拟合能力等优点。本文中神经网络模型考虑了更多临床上的重要病理特征，也突出了其精度更高，模型拟合效果更好的优点，一样能为 ICC 病人临床决策提供重要参考。

4.2. 建议

1) 做好医疗宣传，对肝内胆管癌的隐秘性多加描述，针对肝内胆管癌疾病的特殊性即诊断时多为晚期。发病人群多为老年人，考虑到年龄较大，其预后是极差的，一定要定期体检，早日查出病情，尽早就医能有效延长患者的生存时间。

2) 截止目前，肝内胆管癌的成因依旧是医学界的一大疑问。所以关于 ICC 的预防应主要集中于对其密切相关的疾病及癌前病变的早期治疗。无创伤性检查 B 超应作为该系疾病普查的基本手段。

a) 一级预防：肝内胆管癌病因尚不清楚，与胆结石症的关系也不如胆囊癌密切。因此，胆管癌的一级预防缺乏有效的方法。主要是对肝胆管结石的防治以及定期的系统的健康检查。

b) 二级预防：二级预防是本病预防的重点。阻塞性黄疸患者，在排除胆石症、肝炎、肝硬化等疾病，应高度警惕胆管癌的可能。在详细询问病史、全面体格检查的基础上，应尽早做 B 超、CT、PTC 及 ERCP 检查，以便早期发现、早期诊断、早期治疗。

综上所述，肝内胆管癌是恶性程度较高，治疗较困难，预后较差的一种类型。目前，仅肝切除术对可切除肝内胆管癌病人具有明确延长生存时间作用。肝内胆管癌的治疗方法也会随着技术手段的创新与突破逐渐体现出个体化与选择性。有些仍有争议的细节，还需要大量依据以进一步研究。无论哪种治疗方法均是为了提高患者的生活质量，延长生存期，在不同的情况下要采取不同方法来达到最佳治疗效果。

4.3. 研究展望

本文基于三种预测模型对 ICC 病人预后进行预测，在传统 Cox 风险回归模型进行预测的基础上考虑了加入机器学习算法建立更多模型对比。结果表明基于机器学习的梯度提升法和神经网络模型都能完成精度更高的预测。但由于诸多条件因素的限制以及时间的紧迫性，本项研究存在以下局限性：

1) 本研究为基于 SEER 数据库的回顾性分析，因此本研究建立的模型仍需在外部的大样本、多中心数据集以及前瞻性研究中进一步证实其临床应用价值。尽管如此，此次预测模型中的主要特征，如肿瘤大小、肿瘤数目和区域淋巴结转移数目，均为公认的影响 ICC 预后的因素，体现了本研究建立模型的可靠性。

2) 由于本研究纳入分析的病人和肿瘤特征有限，虽然本研究构建的模型所涵盖的一些协变量极易从临床资料中获得，证明了该模型的简洁性与可行性。但是在缺少其他复杂协变量的情况下，本研究建立的模型并不能够提供较准确的预后预测。

3) 本研究建立的多个模型能够明确一些患者的预后，但无法为具体辅助治疗方案的选择提供依据。

综上所述，本研究构建的模型仍然能为 ICC 病人临床决策提供重要参考。要想更好预测病人的预后，在往后的研究过程中，要充分利用真实病人的数据，改进现有模型算法。另外，机器学习等智能化的手段将逐步取代传统算法，期待其在大数据时代将发挥重要的临床价值。

参考文献

[1] 赵燕青, 董辉, 丛文铭. 肝内胆管癌病理学分型新进展[J]. 肝脏, 2022, 27(2): 242-244. https://doi.org/10.14000/j.cnki.issn.1008-1704.2022.02.015

[2] 沈锋, 王葵. 肝内胆管癌诊断和治疗焦点问题[J]. 中国实用外科杂志, 2020, 40(6): 644-649. https://doi.org/10.19538/j.cjps.issn1005-2208.2020.06.05

[3] 潘晓涛, 南虹, 曹秋祥. 肝内胆管癌患者手术治疗与术后联合放疗的疗效对比: 基于 SEER 数据库的倾向评分匹配研究[J]. 现代肿瘤医学, 2022, 30(3): 495-499.