2015 年全国大学生统计建模大赛

参赛题目：

非小细胞型肺癌分型介导基因 UCN2|90226

对肺癌生存时间影响的研究[1]

参赛学校：南京医科大学

参赛组别：本科生组

参赛选题：大数据统计建模类

参赛队员：朱晓璇 于星灿 殷纪鹏

指导老师：赵杨 魏永越

2015.6

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类本科生组二等奖。

非小细胞型肺癌分型介导基因 UCN2|90226 对生存时间影响的研究

摘 要

目的：肺癌的治疗及预后开始陷入瓶颈，发展日趋缓慢。基因治疗作为目前热门的治疗方法，被广大研究者所认可，并已有所成效。对非临床实验工作者从数据中挖掘有用的基因信息，为肺癌的治疗及预后提供可行的通路，是本文的研究目的。**方法**：本研究拟建立基因-肿瘤分型-非小细胞肺癌患者生存时间的中介效应的生存分析模型，运用了 Logistic 回归、生存分析、Cox 比例风险模型的方法分析数据。**结果**：根据 3 种模型，筛出基因 UCN2|90226，按检验水准 0.05，其对生存时间总影响有统计学意义(P<.0001,HR=1.008)，说明该基因的表达水平每增加一个单位，肿瘤的生存时间越短，死亡风险越高，该基因是个危险因素。**结论**：研究表明，基因 UCN2|90226 对非小细胞型肺癌的生存时间的影响可以通过肿瘤分型作用。基因 UCN2|90226 表达水平越高，越容易患鳞癌，鳞癌会增大死亡的风险。

一、问题产生

(一) 研究背景

肺癌(Lung Cancer)是一个全球性的健康问题，是世界上发病率和死亡率都较高的癌症之一，每年约有 120 万人被诊断为肺癌，而中国的肺癌患病绝对人数占全世界的第一位[1]。根据肿瘤的生物学行为和治疗、预后等因素将肺癌分为非小细胞肺癌(non-small cell lung cancer, NSCLC)和小细胞肺癌(small cell lung cancer, NSCLC)。非小细胞型肺癌包括鳞状细胞癌(鳞癌,adenocarcinoma)、腺癌(squamouscarcinoma)、大细胞癌(large cell lung cancer)，与小细胞癌相比，非小细胞肺癌的癌细胞生长分裂较慢,扩散转移相对较晚。非小细胞肺癌约占所有肺癌的 80%，约 75%的患者发现时已处于中晚期，5 年生存率很低。

肺鳞癌又称肺鳞状上皮细胞癌，包括梭形细胞癌，是最常见的类型，占原发性肺癌的 40%～51%。其多见于老年男性，与吸烟有密切关系。肺鳞癌以中央型肺癌多见，并有胸管腔内生长的倾向，早期常引发支气管狭窄，或阻塞性肺炎。肺鳞癌生长缓慢，转移晚，手术切除机会较多，5 年生存率较高，肺鳞癌对放疗、化疗不如小细胞未分化癌敏感。肺腺癌是肺癌的一种，属于非小细胞癌。不同于

鳞状细胞肺癌 ，肺腺癌较容易发生于女性及不抽烟者。起源于支气管粘膜上皮，少数起源于大支气管的粘液腺。发病率比鳞癌和未分化癌低，发病年龄较小，女性相对多见。多数腺癌起源于较小的支气管，为周围型肺癌。早期一般没有明显的临床症状，往往在胸部 X 线检查时被发现。表现为圆形或椭圆形肿块，一般生长较慢，但有时早期即发生血行转移。淋巴转移则发生较晚。肺大细胞癌在临床上较为罕见，只占全部收治肺癌病例的 1%左右，其恶性程度高，治疗效果差，预后不良，因此肺大细胞癌的治疗关键在于早期发现早期诊断早期治疗。

近几十年来, 我国肺癌的发病率呈逐渐增高趋势, 虽然随着手术技巧的提高, 以及配合放疗、化疗等各种综合手段的应用，肺癌患者的预后有一定改善，但由于肺癌是多种癌基因及抑癌基因改变和多阶段变化所致的独特的生物学行为, 因此对有关肺癌的肿瘤基因进行研究，对预后的判断具有重要意义。

(二) 国内外研究现状

随着分子生物学、免疫学等相关科学的的发展和交叉渗透，基因治疗的研究突飞猛进。目前，已有多种对肺癌预后有重要影响意义的基因为大家所发现。例如，原癌基因中的 K-ras 基因，尽管在非小细胞肺癌早期患者的个体研究中还未达到统计学的显著性，但 K-ras 基因突变仍预示患者预后不良[2]。原癌基因中的 C-myc 基因，其扩增很早就被认为与肺癌有关，尤其是小细胞肺癌中。抑癌基因中的 P16 基因，其在 1992 年被发现，该基因在肺癌的生长发展中起着十分重要的作用。抑癌基因中的 P53 基因可作为肺癌, 尤其是腺癌生存率低的一个重要预测因素[3] 。21 世纪新发现的抗凋亡基因——survivin 基因被认为与胚胎发育及肿瘤发展有关。其作为细胞凋亡因子，对肺癌预后有重要影响[2]。细胞周期调控相关基因 $P27^{KIPI}$ 能够抑制细胞周期依赖性激素, 阻断细胞周期从 G1 到 S 期的转化, 从而抑制了肿瘤的形成。在肺癌中，$P27^{KIPI}$ 与预后关系较为复杂，因为肺癌存在多种类型，但的 $P27^{KIPI}$ 作用不容忽视。除去以上提及的几种具有代表性的基因外，仍有其他一些肺癌相关基因对预后有直接或间接的影响。

(三) 问题提出

寻找对肺癌治疗与预后有意义的基因，并了解这些基因决定肺癌预后的机制，是我们对于生命科学研究的一个切入点。每一种预后相关基因的发现，都为肺癌的治疗提供可能可行的通路，增进肺癌治疗与预后的发展进程。如何发现肺癌治疗与预后的相关基因，是目前的热点问题。

在大数据(Big data)时代，生命科学无疑会成为世界的主角。整个生物组学的大数据已经达到 10 的 60 次方的数量级，而人类现在只完成了 10 的 21 次方，如果没有大数据我们将寸步难行。基因作为生命科学最重要的一部分，一直是广大

生物学研究者的重点目标。

网络上很多与基因组学有关的数据库，如癌症基因组图谱计划(The Cancer Genome Atlas, TCGA)，它是企图全面的并列的去努力地加速理解癌症的分子基础，通过利用包括大规模基因组测序的基因组分析技术来实现。由美国国立癌症研究所(National Cancer Institute, NCI)和美国国立人类基因组研究所(National Human Genome Research Institute, NHGRI)用分阶段的策略来启动；ENCODE，全称“DNA 元素百科全书”计划，该计划获得了迄今最详细的人类基因组分析数据；1000 genomes，全称“全球千人基因组计划”，千人基因组旨在对全球各地的 2500 个人的 DNA 进行比较和排序，将有助于人们进一步了解各种疾病并寻找新的治疗方法。该计划的领导人员宣布他们已成功绘制 95%的人类基因图谱。这些数据每天都在增长，蕴含着丰富的信息。如何充分利用这些在线网络数据库，挖掘数据内隐藏的重点，已成为当前研究的热点。

在肺癌病人疾病进展或死亡的过程中，存在着诸多因素的作用。例如，基因对病人的预后生存时间可能存在着直接的影响[4][5]，肿瘤的病理分型也可能发挥着重要作用[6]。然而，这些因素往往并非单独发挥作用，例如，是否存在着基因不但能直接影响肺癌生存时间，也能通过导致不同的病理分型从而间接地影响生存时间？了解这一点，有助于根据病人相关基因表达水平及病理分型，有针对性地采取治疗方案。

基于以上假设，我们立足于大数据时代，从网络在线数据库下载肺癌基因和临床数据，利用中介效应分析(mediation analysis)，建立相关模型，去探讨寻找是否存在一种或多种基因，它对于病人预后生存时间有影响，同时其也通过影响肿瘤分型继而去影响生存时间。为肿瘤的治疗和预后提供新的思路。

(四) 中介效应分析

中介效应是指变量间的影响关系($X \rightarrow Y$)不是直接的因果链关系，而是通过一个或一个以上变量(M)的间接影响产生的，此时我们称 M 为中介变量(mediator)，而 X 通过 M 对 Y 产生的间接影响称为中介效应。中介效应是间接效应的一种，模型中在只有一个中介变量的情况下，中介效应等于间接效应；当中介变量不止一个的情况下，中介效应不等于间接效应，此时间接效应可以是部分中介效应的和/或所有中介效应的总和。在理论上，中介变量意味着某种内部机制[7]。自变量 X 的变化引起中介变量 M 的变化，中介变量 M 的变化引起因变量 Y 的变化。例如，某种治疗癌症的药物(X)需要通过特定的酶(M)才能有效杀死肿瘤细胞(Y)，如果体内缺少这种酶，药物的作用将失效。

二、模型构建前的准备

(一) 数据来源与描述

实验所需的研究对象均来自 TCGA 数据库，包括研究对象的生存基线数据、肺腺癌和肺鳞癌组织基因表达水平三个数据集。分析数据集中腺癌数据集为来自北美 163 例未经治疗的肺腺癌患者。腺癌的表达水平通过 llumina HiSeq 检测。队列的中位随访时间为 19 个月，72%的患者在上一次随访时存活(2011.11)，81%的患者有吸烟史；肺鳞癌数据集为来自北美 168 例未经治疗的 I-IV 期肺鳞癌患者的肿瘤样本。鳞癌的表达水平通过 Agilent 244k microarrays 和 Hiseq 2000 RNAseq 检测。中位随访时间为 15.8 个月，60%的患者在上一次随访时存活(2011.11)，96%的患者有吸烟史。所有患者均提供书面知情同意进行基因组研究，通过照当地机构审查委员会审查。肺癌病理专家委员会对所有样本进行了额外的检查，以确认组织学亚型。原数据库链接地址：

https://tcga-data.nci.nih.gov/docs/publications/luad_2014/

https://tcga-data.nci.nih.gov/docs/publications/lusc_2012/

(二) 研究指标

确定研究指标：分析数据集研究指标如下表(只包括可供分析的人，即生存时间和生存状态均不缺失的人)

表 1　研究指标一览

变量	标签	取值
Patient	病人(Patient)	TCGA_***_***
GENDER	性别(Gender)	1=Male 0=Female
NPYS	吸烟量(包/年)	连续型变量
SMOKE	吸烟量	0=NPYS≤中位数(40) 1=NPYS>中位数(40)
OUTCOME	结局状态	0=alive 1=deceased
TIME	生存时间	连续型变量
T	二分类时间	0=TIME≤K-M交叉点时间(1220) 1=TIME>K-M交叉点时间(1220)
CENSOR	删失状态	0=删失 1=完全
TYPE	肿瘤分型	0=腺癌 1=鳞癌
STAGE	肿瘤分期	0=早期(一、二期) 1=晚期(三、四期)
基因变量		连续型变量

表 2　研究指标关于肿瘤类型组间比较情况

变量	说明	腺癌(n=163)	鳞癌(n=168)	组间比较
性别	男	76	121	P<0.0001
	女	87	47	
吸烟量	≤40包	95	75	P<0.0131
	>40包	68	93	
肿瘤分期	早期	117	129	P=0.2972
	晚期	46	39	
删失	完全	46	96	P=0.0054
	删失	117	72	
中位生存时间	天	1287	1426	P=0.5364

(三) 假设模型

1. 生存分析模型

常被应用在流行病学和社会科学上进行生存分析的有两种模型：比例风险模型(proportional hazards model)和加速失效时间模型(accelerated failure time model, AFT)。在比例风险模型中 Cox 比例风险模型以半参数方式出现，适用于许多不同或未知分布的资料，协变量还可以是时依协变量(time-dependent covariates)等优势被广泛使用。我们对本研究涉及的生存过程考虑 Cox 比例风险模型和加速失效模型两种模型，选择其一进行生存分析。

(1) Cox 比例风险模型

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{X})$$

(2) 加速失效模型

$$\log(T) = \boldsymbol{\beta}'\mathbf{X} + e$$

(3) 如何选择模型

在实际应用中，我们常常在拟合模型以前，先作 $\log\{-\log S_x(t)\}$ 对 $\log(t)$ 的关系图，以判断哪种模型较为适合[8]，当资料符合比例风险假定时，在同一座标纸上作出的图形显示，两条图线大致平行；而当资料符合 AFT 模型时，一条图线大致可由另一条图线沿 $\log(t)$ 平移得到[9]。如下图：

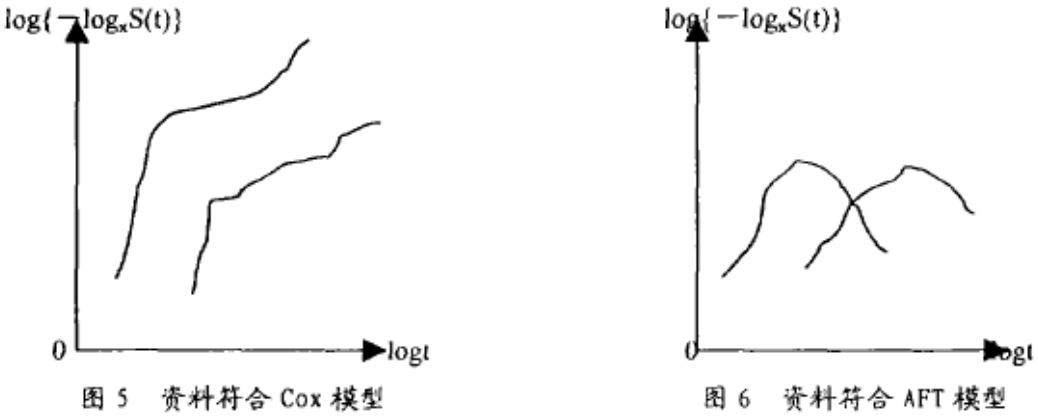

图 5　资料符合 Cox 模型　　图 6　资料符合 AFT 模型

图 1　Cox 比例风险和加速失效模型区别图 (来源：师成虎，2003)

2. 中介效应基本模型

以最简单的中介模型为例说明。考虑自变量 X 对因变量 Y 的影响，如果 X 通过影响 M 而对 Y 产生影响，则称 M 为中介变量。假设所有变量都已经中心化(即将数据减去样本均值，中心化数据的均值为 0)或者标准化(均值为 0，标准差为 1)，可用下列回归方程来描述变量之间的关系[10]，图 1 为相应的路径图：

$$Y = cX + e_1 \qquad \text{公式(1)}$$

$$M = aX + e_2 \qquad \text{公式(2)}$$

$$Y = c'X + bM + e_3 \qquad \text{公式(3)}$$

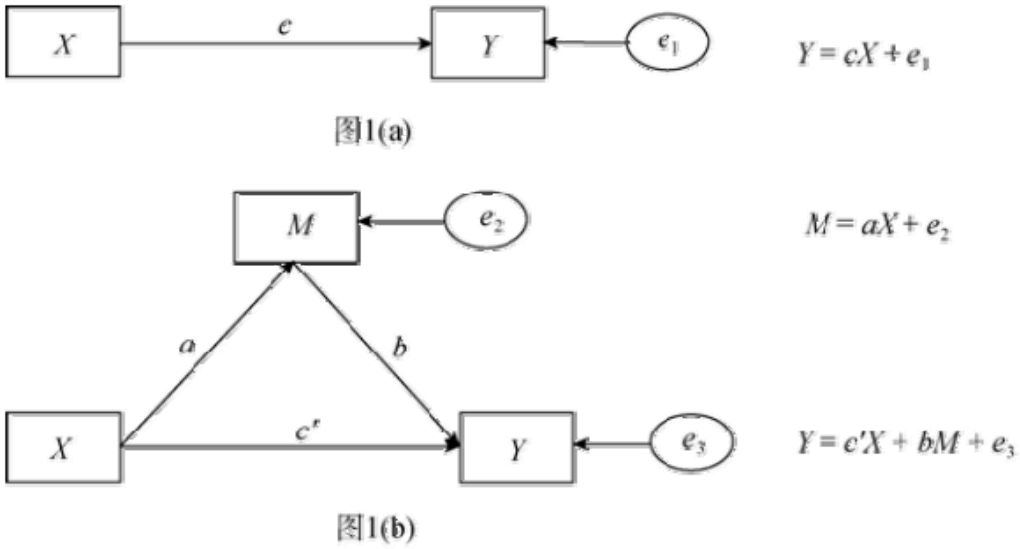

图 2　中介模型示意图 (来源：温忠麟，2008)

其中公式(1)的系数 c 为自变量 X 对因变量 Y 的总效应(Total Effect,TE)；公式(2)的系数 a 为自变量 X 对中介变量 M 的效应，称为自然间接效应(Natural Indirect Effcet,NID)；公式(3)的系数 b 是在控制了自变量 X 的影响后，中介变量 M 对因变量 Y 的效应；系数 c'是在控制了中介变量 M 的影响后，自变量 X 对因变量 Y 的直接效应，称为受限制的直接效应(Controlled Direct Effect,CDE)；e_1~e_3 是回归残差。对于这样的简单中介模型，中介效应等于间接效应(indirect effect)，即等于系数乘积 ab，它与总效应和直接效应有下面关系[12]：

$$c = c' + ab \qquad \text{公式(4)}$$

3. 中介效应的检验方法

检验中介效应最常用的方法是逐步检验回归系数[13]，即通常说的逐步法：

(1) 检验公式(1)的系数 c (即检验 H_0 ：c =0)；

(2) 依次检验公式(2)的系数 a (即检验 H_0 ：a = 0)和公式(3)的系数 b (即检验 H_0 ：b = 0)，有文献称之为联合显著性检验[14]。如果(1)系数 c 显著，(2)系数 a 和 b 都显著，则中介效应显著。

(3) 完全中介过程还要加上公式(3)的系数 c'不显著。

4. 生存分析的中介效应模型

Tein 和 MacKinnon [15][16]给出生存分析的中介效应模型，其中 X 表示暴露，M 表示中介变量，T 表示生存时间，C 表示协变量。

(1) 比例风险的中介效应模型

$$\lambda_T(t \mid X, M, C) = \lambda_T(t \mid 0,0,0)\exp(\beta_1 X + \beta_2 M + \beta_3^{'} C)$$

(2) 加速失效时间中介效应模型

$$\log(T) = \theta_0 + \theta_1 X + \theta_2 M + \theta_4^{'} C + e$$

5. 本研究拟构建的中介效应模型

本研究拟建立基因-肿瘤分型-非小细胞肺癌患者生存时间的中介效应的生存分析模型，如图 3 所示：

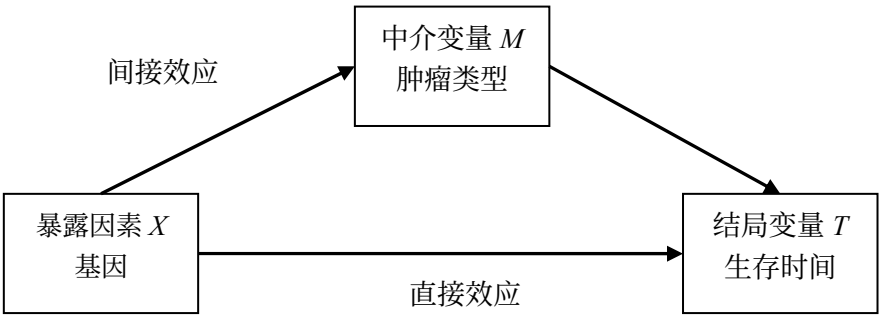

图 3　基因-肿瘤分型-非小细胞肺癌患者生存时间中介效应路径图

三、建模过程

本研究先对肿瘤类型与生存时间的关系进行检验, 并以肿瘤类型分层对生存过程进行描述以选择合适的生存分析模型 ;根据生存分析模型建立中介效应分析中效应分解的三个模型；进而对 19258 个基因标 log2 准化后进行单因素分析, 进行基因筛选；分别按 P<0.005~0.001 筛选出三个模型都符合的基因；最后对特定基因进行标准化后进行中介效应分析与检验。

(一) 生存过程描述

1. 绘制 $\log\{-\log S_x(t)\}$ 对 $\log(t)$ 的关系图，如下图

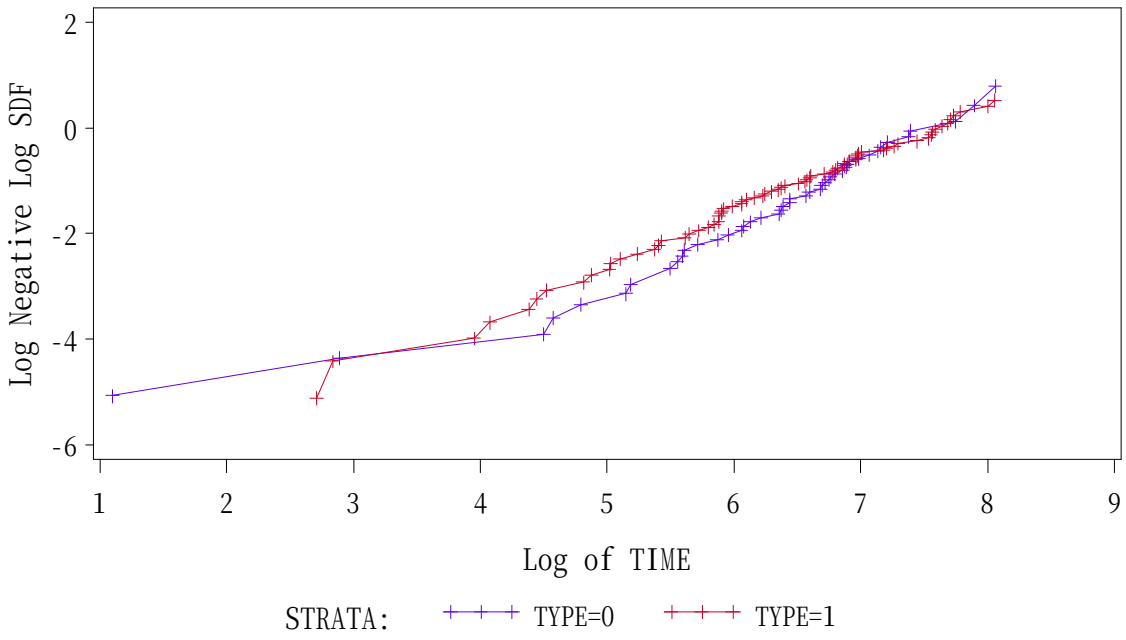

图 4　$\log\{-\log S_x(t)\}$ 对 $\log(t)$ 的关系图

由上图可知，图像平移不明显，似平行又有交叉。初步先考虑 Cox 比例风险模型。

2. 绘制对于不同肿瘤类型绘制不同 Kaplan-Meier 曲线，如下图.

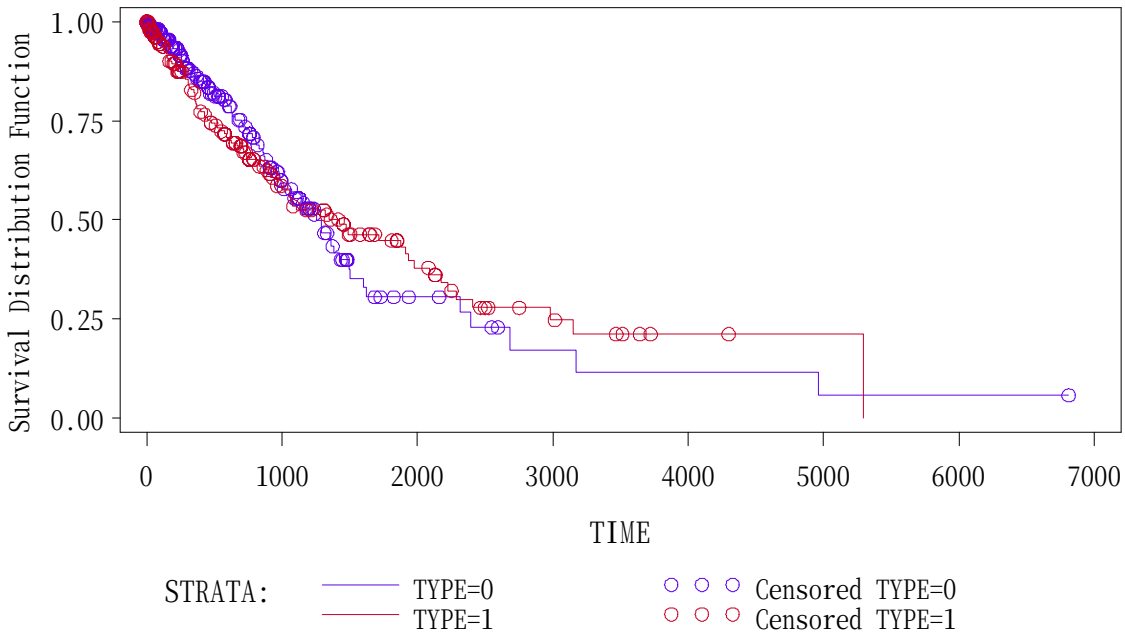

图 5　肿瘤类型 Kaplan-Meier 曲线图

由图 5 可知，生存曲线有相交，不满足等比例风险。

3. 对肿瘤类型和生存时间进行非参数检验，检验结果见下表。

表 3　肿瘤类型和生存时间非参数检验结果

检验方法	卡方统计量	自由度	*P* 值
Log-Rank	0.3822	1	0.5364
Wilcoxon	1.3150	1	0.2515
-2Log(LR)	0.4496	1	0.5025

结论：Log-rank 检验肿瘤类型对生存时间的影响无统计学意义(P=0.5364)。

考虑肿瘤类型的风险比与时间有关，具有时间相依性，是时依协变量，可以考虑分层分析(陶庄，2008)。故将生存时间 TIME 另变为二分类变量 T，对时间 T 分层，分层方法为：从 K-M 曲线交叉处找到 y 轴生存率(近似值范围)，对应到各组 K-M 估计表，取生存率近似相等时对应的时间：1220 天。纳入 Cox 比例风

险模型，其中性别、吸烟量、肿瘤类型以二分类变量作为协变量纳入模型：

$$\lambda(t)=\lambda_{i0}(t)\exp(\beta_2\times TYPE+\beta_3^{'}\times Covariates) \qquad \text{公式(5)}$$

表 4　公式(5)的参数估计结果

变量	估计值	标准误	卡方值	*P* 值	HR(95%CI)
性别	0.1627	0.2007	0.6572	0.4176	1.177(0.794,1.744)
吸烟量	-0.1497	0.1884	0.6314	0.4268	0.861(0.595,1.246)
肿瘤分期	0.54931	0.20008	7.5374	0.006	1.732(1.17,2.564)
肿瘤类型	0.48583	0.19849	5.9906	0.0144	1.626(1.102,2.399)

结论：在时间分层的情况下，肿瘤分型对生存时间的影响有统计学意义。故建立以时间分层的 Cox 比例风险模型。

(二) 模型建立

1. 效应分解

将总影响分解为直接影响和间接影响，影响分解模型如下：

(1) 总影响(TE)：Cox 比例风险模型：

$$\lambda_i(t)=\lambda_{i0}(t)\exp(\beta_{i1}\times GENE_i+\beta_{i3}^{'}\times Covariates), i=1,...,19258. \qquad \text{公式(5)}$$

(2) 自然间接影响(NIE)：Logistic 回归：

$$\text{logit}P(TYPE=1\mid X,C)=\theta_{i0}+\theta_{i1}\times GENE_i+\theta_{i2}^{'}\times Covariates, i=1,...,19258. \quad \text{公式(6)}$$

(3) 受限制的直接影响(CDE)：时间分层的 Cox 比例风险模型：

$$\lambda_i(t)=\lambda_{i0}(t)\exp(\beta_{i1}\times GENE_i+\beta_{i2}\times TYPE+\beta_{i3}^{'}\times Covariates), i=1,...,19258. \quad \text{公式(7)}$$

对上述三个模型进行 19258 个基因的单因素分析，在筛选前，将基因取 log2(GENE)标准化，筛选结果如下：

(1) 三个模型 *P* 值都小于 0.005 的基因的个数：18 个；

(2) 三个模型 *P* 值都小于 0.004 的基因的个数：11 个；

(3) 三个模型 *P* 值都小于 0.003 的基因的个数：8 个；

(4) 三个模型 *P* 值都小于 0.002 的基因的个数：3 个；

(5) 三个模型 *P* 值都小于 0.001 的基因的个数：1 个.

表 5　三个模型 *P* 值都小于 0.005 的基因筛选结果

基因	估计值	标准误	统计量	P 值	HR/OR	效应
ACD|65057	0.4927	0.1727	8.1391	0.0043	1.6368	CDE
	0.5002	0.1674	8.9309	0.0028	1.6491	TE

	1.4740	0.2414	37.2786	0.0000	4.3668	NIE
C20orf197|28475	-0.1304	0.0453	8.2840	0.0040	0.8777	CDE
	-0.1256	0.0441	8.1276	0.0044	0.8819	TE
	-0.3078	0.0592	27.0032	0.0000	0.7351	NIE
CD3EAP|10849	0.4634	0.1592	8.4683	0.0036	1.5895	CDE
	0.4177	0.1343	9.6648	0.0019	1.5184	TE
	1.2898	0.2177	35.1090	0.0000	3.6322	NIE
CSNK1E|1454	0.5119	0.1781	8.2607	0.0041	1.6684	CDE
	0.5357	0.1675	10.2320	0.0014	1.7087	TE
	1.2019	0.2272	27.9931	0.0000	3.3265	NIE
IL11|3589*	0.1563	0.0493	10.0647	0.0015	1.1692	CDE
	0.1743	0.0474	13.5376	0.0002	1.1904	TE
	0.3663	0.0677	29.2729	0.0000	1.4423	NIE
KCNJ14|3770	0.3219	0.0980	10.7896	0.0010	1.3797	CDE
	0.2330	0.0822	8.0395	0.0046	1.2624	TE
	0.8841	0.1350	42.9010	0.0000	2.4208	NIE
LASS4|79603	-0.2520	0.0836	9.0865	0.0026	0.7773	CDE
	-0.2694	0.0833	10.4672	0.0012	0.7638	TE
	0.2910	0.1022	8.1066	0.0044	1.3378	NIE
LDLRAD3|143458	0.2434	0.0857	8.0628	0.0045	1.2756	CDE
	0.2235	0.0710	9.9009	0.0017	1.2504	TE
	1.0759	0.1341	64.3523	0.0000	2.9326	NIE
MAFK|7975	0.2732	0.0902	9.1722	0.0025	1.3141	CDE
	0.2701	0.0922	8.5726	0.0034	1.3101	TE
	-0.8029	0.1496	28.8168	0.0000	0.4481	NIE
MPP7|143098	-0.2428	0.0793	9.3668	0.0022	0.7845	CDE
	-0.2089	0.0688	9.2259	0.0024	0.8114	TE
	-0.8113	0.1293	39.3996	0.0000	0.4443	NIE
PELI2|57161	-0.3305	0.0840	15.4887	0.0001	0.7186	CDE
	-0.3219	0.0804	16.0265	0.0001	0.7247	TE
	0.3169	0.1043	9.2241	0.0024	1.3728	NIE
PHF20|51230	0.5380	0.1799	8.9449	0.0028	1.7126	CDE
	0.5329	0.1813	8.6355	0.0033	1.7038	TE
	-0.7177	0.2341	9.3972	0.0022	0.4879	NIE
RAD21L1|642636	0.5837	0.1826	10.2218	0.0014	1.7926	CDE
	0.4796	0.1568	9.3487	0.0022	1.6153	TE
	-0.7609	0.2395	10.0903	0.0015	0.4673	NIE
TRIB1|10221	0.2857	0.1010	8.0004	0.0047	1.3307	CDE
	0.2828	0.1003	7.9516	0.0048	1.3269	TE
	-1.2252	0.1732	50.0140	0.0000	0.2937	NIE
UCN2|90226*#	0.2071	0.0562	13.5924	0.0002	1.2301	CDE
	0.1444	0.0437	10.9277	0.0009	1.1553	TE
	0.6502	0.0745	76.2376	0.0000	1.9160	NIE
WDR3|10885	0.5358	0.1737	9.5090	0.0020	1.7088	CDE

	0.4432	0.1476	9.0162	0.0027	1.5577	TE
	1.2801	0.2298	31.0259	0.0000	3.5971	NIE
ZC4H2|55906	0.2354	0.0772	9.3048	0.0023	1.2654	CDE
	0.2608	0.0769	11.5021	0.0007	1.2980	TE
	-0.4690	0.1037	20.4606	0.0000	0.6256	NIE
ZNF598|90850*	0.5919	0.1635	13.0996	0.0003	1.8074	CDE
	0.5621	0.1630	11.8971	0.0006	1.7543	TE
	0.6605	0.2058	10.3044	0.0013	1.9358	NIE

备注： *：P 均<0.002；#：P 均<0.001; CDE: 受限制的直接效应(Controlled Direct Effect); TE: 总效应(Total Effect); NIE: 自然间接效应(Natural Indirect Effcet).

表 6　三个模型 *P* 值都小于 0.001 的基因筛选结果

基因	估计值	标准误	统计量	P 值	HR/OR	效应
UCN2|90226	0.2071	0.0562	13.5924	0.0002	1.2301	CDE
	0.1444	0.0437	10.9277	0.0009	1.1553	TE
	0.6502	0.0745	76.2376	0.0000	1.9160	NIE

可知基因 UCN2|90226 统计学意义较大，对其进行中介效应分析。为了增加发现潜在有价值基因的可能性，我们也对三个模型 *P* 均小于 0.002 的基因中另外两个基因：IL11|3589 与 ZNF598|90850 进行了中介分析。

2. 中介效应分析

(1) 以 UCN2|90226 为例，首先对基因标准化：

$$\mathrm{UCN2}\,|\,90226=\frac{\mathrm{UCN2}\,|\,90226_{\mathrm{old}}-mean}{STD} \qquad \text{公式(5)}$$

(2) 给出模型假设

总效应：Cox 比例风险模型：

$$\lambda(t)=\lambda_0(t)\exp(\beta_1\times\mathrm{UCN2}\,|\,90226+\beta_3^{'}\times Covariates) \qquad \text{公式(6)}$$

间接效应：Logistic 回归：

$$\mathrm{logit}P(TYPE=1\,|\,X,C)=\theta_0+\theta_1\times\mathrm{UCN2}\,|\,90226+\theta_2^{'}\times Covariates \qquad \text{公式(7)}$$

受控制的直接效应：Cox 比例风险模型：

$$\lambda(t)=\lambda_0(t)\exp(\beta_1\times\mathrm{UCN2}\,|\,90226+\beta_2\times TYPE+\beta_3^{'}\times Covariates) \qquad \text{公式(8)}$$

肿瘤类型与生存时间的关系：Cox 比例风险模型：

$$\lambda(t)=\lambda_0(t)\exp(\beta_2\times TYPE+\beta_3^{'}\times Covariates) \qquad \text{公式(9)}$$

3. 模型求解

受限制的直接效应、间接效应、总效应、肿瘤类型与生存时间关系的 Cox

比例风险模型，变量的检验均为 Wald 检验，模型的检验均采用似然比检验，检验水准均为$\alpha = 0.05$，模型拟合效果采用赤池信息准则(AIC)。见下表。

表 7　对基因 UCN2|90226 中介效应建模输出结果

模型	变量	系数	卡方值	P值	HR/OR(95%CI)
总效应模型	GENDER	0.22143	1.2229	0.2688	1.248(0.843,1.848)
	SMOKE	-0.16362	0.7531	0.3855	0.849(0.587,1.229)
	STAGE	0.59937	9.1224	0.0025	1.821(1.234,2.6870
	UCN2|90226	0.34438	22.8794	<.0001	1.411(1.225,1.625)
	模型效果：似然比检验(卡方=27.2017，P<.0001)；AIC=1091.097				
受限制的直接效应模型	GENDER	0.22635	1.2435	0.2648	1.254(0.842,1.867)
	SMOKE	-0.13982	0.5468	0.4596	0.870(0.600,1.260)
	STAGE	0.54787	7.4524	0.0063	1.730(1.167,2.563)
	TYPE	0.21764	1.0216	0.3121	1.243(0.815,1.896)
	UCN2|90226	0.29702	15.8226	<.0001	1.346(1.163,1.558)
	模型效果：似然比检验(卡方=25.8201，P<.0001)；AIC=976.647				
间接效应模型	Intercept	0.2039	1.2837	0.2572	.
	GENDER	-0.4773	12.2623	0.0005	0.385(0.226,0.657)
	SMOKE	-0.2617	3.9142	0.0479	0.592(0.353,0.995)
	STAGE	0.1483	0.9384	0.3327	1.345(0.738,2.452)
	UCN2|90226	2.0096	45.6298	<.0001	7.460(4.164,13.366)
	模型效果：似然比检验(卡方=116.3562，P<.0001)；AIC=352.432				
肿瘤类型与生存时间的关系模型	GENDER	0.16270	0.6572	0.4176	1.177(0.794,1.744)
	SMOKE	-0.14970	0.6314	0.4268	0.861(0.595,1.246)
	STAGE	0.54931	7.5374	0.0060	1.732(1.170,2.564)
	TYPE	0.48583	5.9906	0.0144	1.626(1.102,2.399)
	模型效果：似然比检验(卡方=13.9639，P=0.0074)；AIC=986.503				

(三) 模型检验

1. 对基因 UCN2|90226 中介效应的检验

(1) 检验公式(6)的系数β_1(即检验 $H_0:\beta_1=0$), P<.0001, 在检验水准$\alpha = 0.05$下有统计学意义；

(2) 依次检验公式(7)的系数θ_1(即检验 $H_0:\theta_1=0$)和公式(8)的系数β_2(即检验 $H_0:\beta_2=0$)。(7)中系数θ_1的 P<.0001, (8)中系数β_2的 P<.0001, 在检验水准$\alpha = 0.05$下，θ_1与β_2均有统计学意义，则中介效应有统计学意义。

(3) 进一步检验公式(8)的系数β_1，P=0.3121，在检验水准$\alpha = 0.05$下无统计

学意义，说明可能存在其他中介，为部分中介效应。

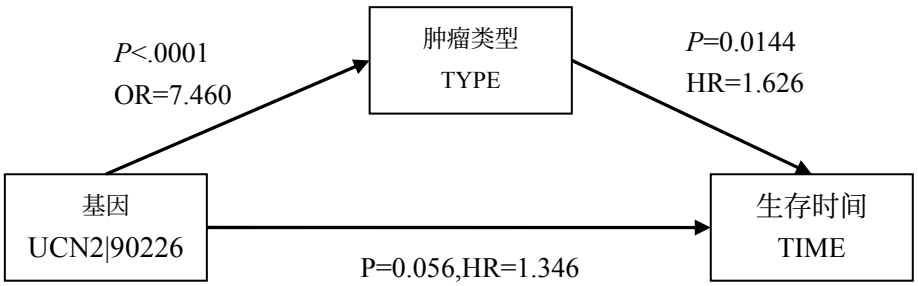

图 6 基因 UCN2|90226 中介效应图

2. 对基因 IL11|3589 与 ZNF598|90850 中介效应检验

基因 IL11|3589 的中介效应有统计学意义, 但为部分中介效应(公式(8)中 β_1 的 P=0.1031)；基因 ZNF598|90850 中介效应有统计学意义, 同样为部分中介效应(公式(8)中 β_1 的 P=0.0560)。

三种基因的中介效应检验结果见下表。

表 8 三种基因中介效应检验结果

基因	NIE		TYPE 与 TIME 关系		CDE		TE	
	P 值	OR(95%CI)	*P* 值	HR(95%CI)	*P* 值	HR(95%CI)	*P* 值	HR(95%CI)
UCN2|90226	<.0001	7.460(4.164,13.366)	0.0144	1.626(1.102,2.399)	<.0001	1.346(1.163,1.558)	<.0001	1.411(1.225,1.625)
IL11|3589	<.0001	1.411(1.225,1.625)	0.0144	1.626(1.102,2.399)	0.0056	1.249(1.067,1.462)	<.0001	1.350(1.166,1.564)
ZNF598|90850	0.0002	1.648(1.266,2.145)	0.0144	1.626(1.102,2.399)	0.0007	1.323(1.126,1.555)	0.0008	1.328(1.124,1.569)

(四) 结果

对于基因 UCN2|90226，按检验水准 0.05，其对生存时间总影响有统计学意义(P<.0001,HR=1.411)，说明该基因的表达水平每增加一个单位，肿瘤的生存时间越短，死亡风险越高，该基因是个危险因素。且基因对肿瘤类型的影响有统计学意义(P<.0001,OR=7.460)，可以认为该基因的表达水平每增加一个单位，患鳞癌的风险是患腺癌风险的 7.46 倍，该基因表达更有可能患鳞癌；本数据分析得肿瘤类型和生存时间的影响有统计学意义(P=0.0144，HR=1.626)，且鳞癌的风险大于腺癌。中介效应检验有统计学意义，效应分解的结果也能够得到解释，中介效应模型建立成功。可以认为，基因 UCN2|90226 通过影响肿瘤类型，影响生存时间。但该中介效应为部分中介效应，说明还有其他的中介发挥作用。

对于基因 IL11|3589，按检验水准 0.05，其对生存时间总影响有统计学意义(*P*<.0001,HR=1.350)，为危险因素；基因对肿瘤类型的影响有统计学意义(*P*<.0001,OR=1.411)，该基因表达更容易患鳞癌；该中介效应检验有统计学意义且为部分中介效应。

对于基因 ZNF598|90850，其对生存时间总影响有统计学意义(*P*=0.0008,HR=1.328)，该基因是个危险因素。其基因对肿瘤类型的影响有统计学意义(*P*=0.0002，OR=1.648)，该基因表达更容易患鳞癌，其中介效应检验有统计学意义且为部分中介效应。

由于基因 UCN2|90226 对肿瘤类型的影响较大，OR=7.46，故相比于另外两个基因把肿瘤类型为中介的影响意义更大。

四、结论和建议

基因治疗作为非小细胞肺癌研究的主流方向之一，需要越来越多的人去发现其中蕴含的信息。现代公共数据平台的建立，使非临床研究人员的数据挖掘更加便利。本文章借鉴中介效应分析的思想，运用统计学上的 Logistic 回归、生存分析、Cox 比例风险模型等方法，分析 TCGA 数据平台搜集的腺癌、鳞癌病人基因数据，最终发现 UCN2|90226 基因对肺癌肿瘤分型、肺癌病人的生存时间在统计学上有意义。基因 UCN2|90226 在肺癌肿瘤的分型中扮演着重要的角色，其主要是通过调整肺癌肿瘤的病理分型，继而对肺癌病人的预后生存时间产生影响。基因 UCN2|90226 对非小细胞型肺癌的生存时间的影响可以通过肿瘤分型作用。基因 UCN2|90226 表达水平越高，越容易患鳞癌，鳞癌会增大死亡的风险。

UCN2|90226 基因位于 3 号染色体短臂，通过对 genecards 网站检索发现，该基因被认为与促肾上腺皮质激素释放有关。在脑组织中，该基因的表达水平与压力下食欲水平有关。最近研究发现，促肾上腺皮质激素释放可以调节细胞凋亡，从而影响肿瘤细胞的迁移[17]，提示可能与本研究肺癌预后相关。同时，研究表明，促肾上腺皮质激素释放相关基因也与胚胎肺生长有关[17]，提示 UCN2|90226 可能参与了肺癌病理分型的形成。关于 UCN2|90226 基因参与肺癌预后的过程，本研究将在后继分析中，申请 DNA 资料进行进一步分析。

IL11|3589 基因位于 19 号染色体长臂，通过对 genecards 网站检索发现，这种细胞因子可以刺激 B 细胞产生免疫球蛋白的 T 细胞依赖性发展。它被发现支持造血干细胞和巨核系祖细胞增殖。查阅文献发现，该基因和胰腺癌[19]、肝癌[20]等有相关性，当前并没有研究表明该基因与肺癌相关。

ZNF598|90850 基因位于 16 号染色体短臂，通过对 genecards 网站检索发现，

该基因为锌指蛋白结合核酸，在各种细胞功能中起重要作用，包括细胞增殖、分化和凋亡。研究表明该基因可能影响正常和致癌信号网[21]。

中介效应分析在许多领域都有广泛应用，因为它可以分析变量之间影响的过程和机制，相对于回归分析，可以得到比较深入的结果。虽然中介分析不能肯定地说“证实”了什么，但可以帮助我们支持某种理论而排除其竞争的理论。在理论上，中介变量意味着某种内部机制(MacKinnon, 2008)。自变量 *X* 的变化引起中介变量 *M* 的变化，中介变量 *M* 的变化引起因变量 *Y* 的变化。例如，某种治疗癌症的药物(*X*)需要通过特定的酶(*M*)才能有效杀死肿瘤细胞(*Y*)，如果体内缺少这种酶，药物的作用将失效。可见中介变量是参与整个因果过程中的重要一环，不可或缺，正因为如此，中介效应分析的前提是变量间存在明确的(理论上或事实上的)因果关系，否则结果很难解释。

在大数据时代，虽然说更强调相关关系，而对因果关系的追求不再那么执着，但是对于生命科学领域，对因果关系的探讨仍旧至关重要。就本研究而言，研究基因与肿瘤生存时间的中介效应，有助于为临床工作者通过中介效应层面找寻肺癌治疗及预后的新思路。

对于本研究，尽管基因数据大，但样本量较小；没有获得基因分型数据导致牺牲信息；通过查阅文献可知，年龄对肿瘤分型有影响，但数据里没有年龄作为协变量纳入模型，我们将继续努力去获得更为全面的数据，进行分析，做出更有价值的研究。

参考文献

[1] 葛辰蕾.非小细胞肺癌的预后因素分析[D]. 郑州:郑州大学,2011.1-2

[2] 谢晓宇. 肿瘤基因对肺癌预后影响的进展[J].上海第二医科大学学报,2002,22(4):369-371

[3] 曹雪涛.肿瘤基因治疗研究的现状和展望[J].国外医学肿瘤学分册.1996,23(1):1-4

[4] Hun charek M, Kupelnick B, Geschwind JF, et al. Prognostic significance of p53 mutations in non- small cell lung cancer: a meta analysis of 829 cases from eight published studies [J] . Cancer lett , 2000, 153(1-2) : 219- 226.

[5] Rosell R, Li S, Ant on A, et al. Prognostic value of k-ras genotypes in patients with advancer non-small cell lung cancer receiving carboplatin with either int ravenous or chronic oral dose etoposide[J].Int J Oncol, 1994, 5: 169- 176.

[6] 焦霞,印洪林,陆珍凤等.胸腺肿瘤 108 例的病理组织学分型和预后相关性研究[J].中华病理学杂志,2008,37(7)

[7] Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. Organizational Research Methods, 11, 241–269.

[8] Collett,D.ModellingSurvivalDatainMedicalReseareh,Chapman&Hall(1994)

[9] 师成虎.加速失效模型及其在医学研究中的应用[D].山西:山西医科大学,2003.30-31

[10] 温忠麟, 刘红云, 侯杰泰. (2012). 调节效应和中介效应分析. 北京: 教育科学出版社

[11] 温忠麟,叶宝娟.中介效应分析[J].心理学进展,2014,22(5):731-743

[12] MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures.Multivariate Behavioral Research, 30, 41–62.

[13] Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173–1182

[14] Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication Monographs, 76, 408–420.

[15] Tein,J-Y.;MacKinnon,DP.Estimating mediated effects with survival data.In:Yanai,H.;Rikkyo,AO.;Shigemasu,K.;Kano,Y.;Meulman,JJ.,editor.New Developments on Psychometrics.Springer-Verlag Tokyo Inc;Tokyou,Japan:2003.p.405-412

[16] Tyler J.VanderWeele.Causal mediation analysis with survival data[J].Epidemiology.2011 July,22(4):582-585

[17] Jin L.Corticotropin-releasing hormone receptors mediate apoptosis via cytosolic calcium-dependent phospholipase $A_2$and migration in prostate cancer cell RM-1.J Mol Endocrinol. 2014 Apr 28;52(3):255-67.

[18] Chouridou E.A complete corticotropin releasing factor system localized in human fetal lung.Hormones (Athens). 2014 Apr-Jun;13(2):229-43

[19] Ren C.Plasma interleukin-11 (IL-11) levels have diagnostic and prognostic roles in patients with pancreatic cancer.Tumour Biol. 2014 Nov;35(11):11467-72

[20] Nie XH.Comparison of the effects of the pretreatment and treatment with RhIL-11 on acute liver failure induced by D-galactosamine.Eur Rev Med Pharmacol Sci. 2014;18(8):1142-50

[21] Rush J.Immunoaffinity profiling of tyrosine phosphorylation in cancer cells.Nat Biotechnol. 2005 Jan;23(1):94-101. Epub 2004 Dec 12

附录 所用软件和数据集整理过程

一、所用软件为 SAS 9.2

二、数据集整理过程

1.初步整理数据集

做法：

(1) 在 TCGA 网站上调整为统一格式的 Patient 值，即统一识别变量；

(2) 转置生存基线数据集，以 Patient 变量作为识别变量；

(3) 对 LUAD 数据集和 LUSC 数据集纵向拼接，分别对拼接数据集和生存基线数据集按 Patient 变量排序后，进行横向合并：生存基线数据集-鳞癌/腺癌拼接后数据集；

(4) 初步删除不参与分析的观测：

① 对于以下三个变量涉及生存时间，三者取值并联符合以下情况者删除该记录：

情况	Days_to_Death	Days_to_Last_Known_Alive	Days_to_Last_Followup
1	NA	NA	.
2	NA	0	0
3	0	0	0
4	0	NA	.
5	NA	NA	NA
6	0	NA	NA

② 重要变量如肿瘤分型有缺失值的情况。

初步整理后的数据集的情况是：将该数据集共有 480 个观测(即病人)和 19273 个变量(基因变量 19258 个)。其中病人的肿瘤分型为腺癌的数目是 276，为鳞癌的数目是 204。

表 9 第一次整理后变量名列表

类别	变量名称	中文	注释
基本信息	Patient	病人	
	Gender	性别	Male=男 Female=女
	Race	种族	Asian； Black_or_African_American； White； American_indian_or_alaska_native NA
	Cohort	来自数据库	TCGA

生存数据	Vital_Status	生存状态	Alive=尚存 Deceased=已故
	Days_to_Death	生存时间	从诊断到死亡的时间(天数)
	Days_to_Last_Known_Alive	截尾时间	从诊断到上一次生存的时间(天数)
	Days_to_Last_Followup	随访时间	从诊断到最后一次随访时间(天数)
抽烟情况	Number_Pack_Years_Smoked	抽烟量	平均一年抽多少包
	Tobacco_Smoking_History	吸烟史	current_reformed_smoker_for_<_or_=_15_years：过去抽烟，现在不抽，烟龄<=15年； current_reformed_smoker_for_>_15_years：过去抽烟，现在不抽，烟龄>15年； current_smoker：现在仍在抽； lifelong_non_smoker：从不抽烟； NA
	Stopped_Smoking_Year	停止抽烟的时间	年份
诊断数据	Clinicalcqcf_Tumor_Type		Primary=初期
	Year_of_Initial_Pathologic_Diagn	首次病例诊断的时间	年份
	Diagnosis	肿瘤分型	Lung_Adenocarcinoma=肺腺癌 Lung_Squamous_Cell_Carcinoma=肺鳞癌
	Tumor_Stage	肿瘤分期	stage_i=Ⅰ期 stage_ia=ⅠA期 stage_ib=ⅠB期 stage_ii=Ⅱ期 stage_iia=ⅡA期 stage_iib=ⅡB期 stage_iiia=ⅢA期 stage_iiib=ⅢB期 stage_iii=Ⅲ期 stage_iv=Ⅳ期 NA Ⅰ、Ⅱ期为早期，Ⅲ、Ⅳ期为晚期
基因数据	19258条		表达水平

2. 对初步整理后数据集进行再一次整理，用作分析数据集：

① 变量调整：

如果变量 Days_to_Death 已知，该时间值纳入变量 TIME，并令变量

CENSOR=1；如果 Days_to_Death 未知且 Days_to_Last_Followup 已知，将 Days_to_Last_Followup 变量值中时间值纳入 TIME，且 CENSOR=0. (因该数据 Days_to_Last_Followup 已知，Days_to_Last_Known_Alive 必已知，故不考虑 Days_to_Last_Known_Alive 变量。)

对于 Tumor_Stage 删除缺失值，并如果肿瘤分期为一、二期，则令变量 STAGE=0；如果为三、四期，则令变量 STAGE=0。

② 删除不分析变量：

缺失值太多：Race(缺失比例=17.08%)，Tobacco_Smoking_History(缺失比例=93.33%),

对本分析不起作用：Cohort，Stopped_Smoking_Year，Year_of_Initial_Pathologic_Diagn,

变量已转化为新变量(信息量基本不变)：Days_to_Death，Days_to_Last_Known_Alive，Days_to_Last_Followup，Tumor_Stage

提供信息量不全：Clinicalcqcf_Tumor_Type

③ 整理结果：个体：441 人，变量总数 19266，其中基因变量为 19258。

表 10　最终整理后变量列表

变量	**标签**	**取值**
Patient	病人(Patient)	TCGA_***_***
GENDER	性别(Gender)	1=Male 0=Female
NPYS	抽烟量(包/年) (Number_Pack_Years_Smoked)	连续型变量
OUTCOME	结局状态(OUTCOME)	0=alive 1=deceased
TIME	生存时间 (TIME)	连续型变量
CENSOR	删失状态 (CENCOR)	0=删失 1=完全
TYPE	肿瘤分型(TYPE)	0=腺癌 1=鳞癌
STAGE	肿瘤分期(STAGE)	0=早期(一、二期) 1=晚期(三、四期)
基因变量		连续型变量

表 11　变量构成情况描述

变量	频数	百分比%
TYPE		
0=腺癌	239	54.2
1=鳞癌	202	45.8

STAGE		
0=早期	330	74.83
1=晚期	111	25.17
GENDER		
0=女	185	41.95
1=男	256	58.05
OUTCOME		
0=生存	281	63.72
1=死亡	160	36.28
CENSOR		
1=完全	160	36.28
0=删失	281	63.72