文章编号：1004-9231（2023）08-0831-09

· 方法学拾粹 ·

医学研究中常见统计分析方法的应用

曾慧娴[1]，杨之雨[1]，柳东红[2]，王瑞华[1]，陈宏森[3]，张宏伟[3]，谭晓契[3]，李萍[3]，曹广文[1,3]

1. 暨南大学基础医学院流行病学教研室，广东 广州 510632；2. 海军军医大学第三附属医院肝胆外科，上海 200433；
3. 海军军医大学流行病学教研室，上海 200433

摘要：

统计学在医学研究中起着重要的作用，选择合适的统计学方法对于医学研究能否得出可靠和有价值的结论至关重要。本文简要介绍常用的医学数据统计分析方法，包括描述性分析、参数检验、非参数检验、相关分析、回归分析和生存分析。重点讨论多重线性回归、logistic 回归、Cox 比例风险回归的假设以及针对不同的研究目标和数据类型来选择合适的统计方法来分析和解释医学数据。

关键词： 描述性统计方法；参数检验；非参数检验；相关分析；回归分析；生存分析

中图分类号： R181.2 **文献标志码：** A **DOI：** 10.19428/j.cnki.sjpm.2023.22771

引用格式： 曾慧娴，杨之雨，柳东红，等.医学研究中常见统计分析方法的应用[J].上海预防医学，2023，35（8）：831-839.

Selection and application of statistical methods in medical research

ZENG Huixian[1], YANG Zhiyu[1], LIU Donghong[2], WANG Ruihua[1], CHEN Hongsen[3], ZHANG Hongwei[3], TAN Xiaojie[3], LI Ping[3], CAO Guangwen[1,3]

1. *Department of Epidemiology, School of Basic Medical Science, Jinan University, Guangzhou, Guangdong* 510632, *China*;
2. *Department of Hepatic Surgery, the 3rd Hospital Affilliated to Naval Medical University, Shanghai* 200433, *China*;
3. *Department of Epidemiology, Naval Medical University, Shanghai* 200433, *China*

Abstract: Statistics plays an important role in medical research, and the selection of appropriate statistical methods is crucial for drawing reliable and valuable conclusions. This paper provides a brief introduction to commonly used statistical analysis methods for medical data, covering descriptive analysis, parametric test, nonparametric test, correlation analysis, regression analysis, and analysis of survival data. It focuses on discussing the assumptions of multiple linear regression, logistic regression and Cox proportional risk regression, as well as how to choose the appropriate statistical methods for analyzing and interpreting medical data based on different research objectives and data types.

Keywords: descriptive statistics; parametric test; nonparametric test; correlation analysis; regression analysis; survival analysis

医学统计是一门由医学和统计学组成的交叉学科，应用统计学原理和方法来收集、处理、分析和推断医学数据。自统计学引入医学研究以来，它已经成为医学研究中一个重要的工具。在统计学未引入医学研究之前，医学工作者往往无法正确解释自己的结果、结论和主张，也无法正确评价他人的医学研究。

医学数据有其特殊性，一方面医学数据种类繁多，数量庞大，数据关系复杂；另一方面，各种医疗机构临床检测指标的换算单位可能不一样。此外，由于各地区的医疗条件不同，还有相当一部分数据仍然是通过手工录入来收集的。因此，原始医学数据不可避免地包含一些不正确或矛盾的数据，如果不加以解决，将对统计结果产生不可预知的影响。因此，医学统计前的数据清洗工作十分必要。

选择合适的统计方法也是医学数据分析过程中非常重要的一步，确保了结果的可靠性。选择错误的统计方法会在解释研究结果时产生一些严重的后果，甚至还会影响研究的结论。在统计学中，任何统计分析方法都有其应用原则与适用条件。为了选择适当的统计方法，不仅需要了解收集的数据类型和研究目的，还要了解统计方法的适用条件，这样才能选择合适的统计方法进行数据分析[1]。随着计算机技术发展，借助 SPSS、Stata、SAS 和 R 等统计软件可以很容易地进行统计分析。然而，选择适当的统计方法对于医学研究人员而言仍是一项困难的任务，尤其对于没有统计学背景的人群。在已发表的医学研究文章中，使用不恰当或错误的统计方法是一个普遍现象，例如误用 t 检验和 χ^2 检验、不适当地使用参数检验等[2]。

医学统计分析方法种类较多，但医学研究经常使用的方法包括描述性统计、参数检验（t 检验、方差分析 ANOVA 等）、非参数检验（χ^2 检验、Wilcoxon 秩和检验、Kruskall-Wallis 秩和检验等）、相关分析、回归分析（线

【基金项目】上海市公共卫生三年行动计划（GWV-10.1-XK17）

【作者简介】曾慧娴，女，硕士在读；研究方向：公共卫生；E-mail：hxzeng@stu2021.jnu.edu.cn。杨之雨，女，硕士在读；研究方向：公共卫生；E-mail：yangzhiyu@stu2021.jnu.edu.cn。并列第一作者

【通信作者】曹广文，E-mail：gcao@smmu.edu.cn

性回归、logistic回归、Cox回归和Poisson回归）和生存分析[3-5]。本文就以上分析方法逐一介绍，侧重于为特定的医学问题选择正确的统计方法。医学数据的统计分析步骤见图1。

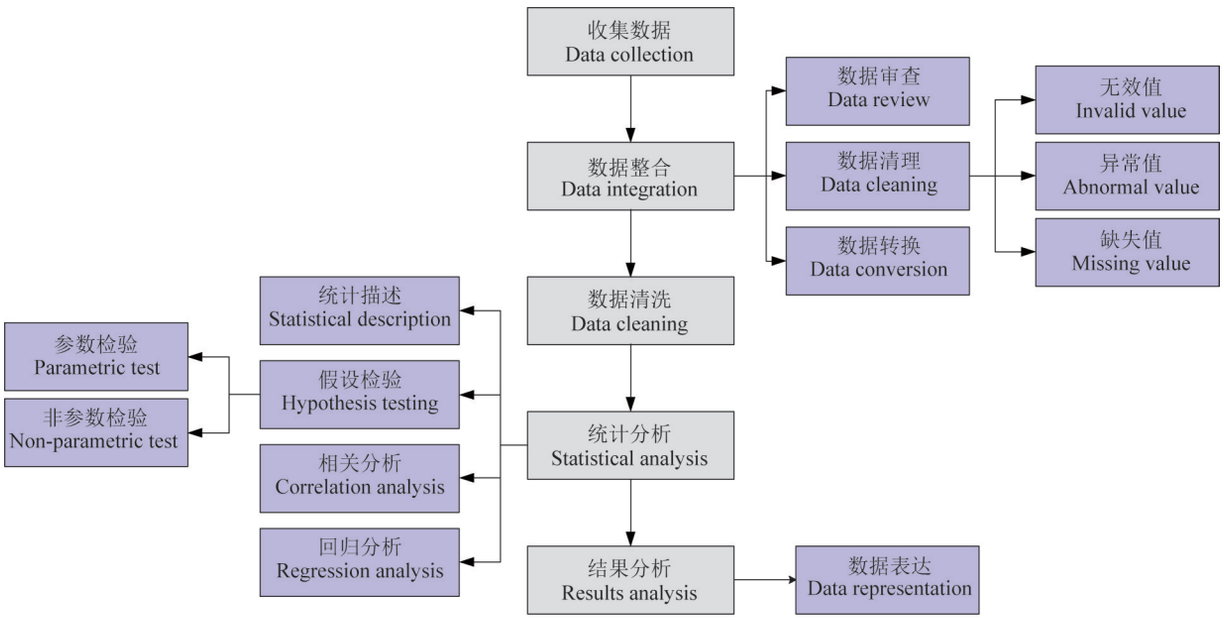

图1　医学数据的统计分析步骤

Figure 1　Steps for statistical analysis of medical data

1　数据清洗

由于受到主观因素、实验环境、不可抗力等因素的影响，原始临床数据中不可避免地会出现一些问题甚至错误。因此在数据审查过程中，需要对发现的问题进行分类汇总。数据清理主要包括3个部分：无效值、异常值和缺失值的处理。对于无效数据，通常会做删除处理。如果能够找到数据来源，并得到原始的正确数据，可以对数据进行重新校正。

异常值的处理方法包括删除、平均值修正等。发现异常值是处理异常值的难点，可通过正态分布3σ原则和画箱图进行异常值的识别。正态分布3σ原则是指在正态或近似正态分布的样本中，一般可以认为数据Y的取值几乎全部集中在$(\mu-3\sigma,\mu+3\sigma)$内（其中$\mu$为平均值，$\sigma$为标准差），超出这个范围的可能性<0.3%，可以认为是异常值。

缺失值是最常见的数据问题，可以删除或忽略，也可以用适当的估计值插补。在数据较少时应考虑使用数据插补将缺失的数据补齐，如多重插补法、K-最近邻法、线性回归法等[6-7]。当缺失率<50%时，插补效果较好；当缺失率增大至50%时，插补效果变差[8-9]。此外，可以根据研究需要将连续变量转换成分类变量，以便深入分析研究因素与结局事件的相关性。将连续变量转换为分类变量的常用方法有参考值范围、指南、共识、均值、中位数和百分位数等[10]。例如，世界卫生组织（WHO）使用体重指数（BMI）来将一个人定义为偏瘦（<18.5 kg·m^{-2}）、正常（18.5~24.0 kg·m^{-2}）、超重（24.1~28.0 kg·m^{-2}）或肥胖（>28 kg·m^{-2}）。没有经过预处理的数据可能会影响后续的统计分析工作。因此，应该仔细研究医学数据的特点，然后对原始数据进行清洗。

2　描述性统计

完整的数据集可能包括成千上万的观察值，但在发表的论文中不可能写出全部结果，因此总结这些数据集信息需要描述性统计[11]。统计描述须结合数据的类型和性质进行。在医学研究中，常见的资料类型有2种，包括定性资料和定量资料。见图2。

对于定性资料，一般用频数和百分比描述。对于定量资料，首先进行正态性检验，如果变量近似正态分布，一般用平均值（标准差）或$\bar{x}\pm s$来描述；如果变量呈现偏态分布，一般用中位数$M(Q_1\sim Q_3)$来描述。正态性检验主要用于评估连续型变量是否服从或近似服从正态分布，常用的方法有Shapiro-Wilk检验、Kolmogorov-Smirnov检验、直方图、P-P图和Q-Q图等[12]。当数据呈非正态分布时，可以通过对数、倒数或平方根转换的方式将数据转化为正态分布或近似正态分布[13]。

医学研究报告中的第一个描述性表格通常用于描述研究对象的基线特征（如人口学、临床和社会等方

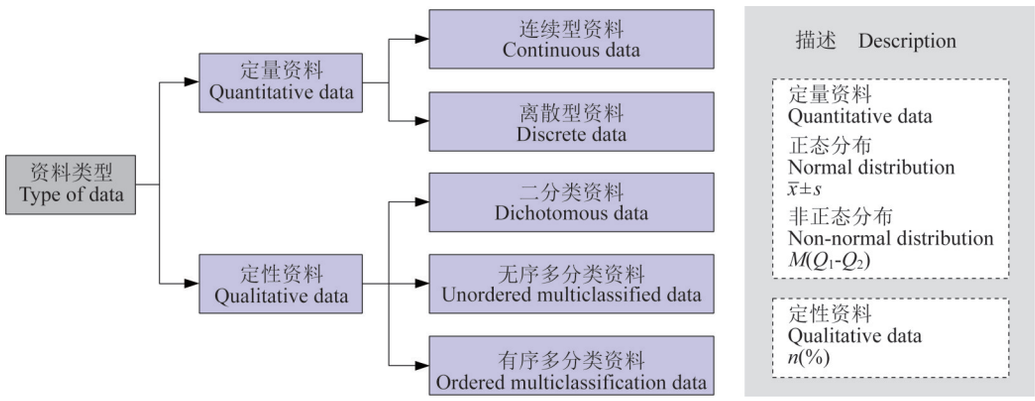

图2 常见的统计资料类型

Figure 2 Common types of statistical data

面)以及关于暴露和潜在混杂因素的信息[14]。表中参与者的性别、种族、糖尿病患病情况和是否吸烟属于定性资料,用频数和百分比描述;而年龄、BMI和基线25-羟基维生素D水平属于定量资料且分布近似正态分布,用$\bar{x}\pm s$来描述。描述性统计的一般表格形式见表1[15]。

表1 参与者的基线特点

Table 1 Baseline characteristics of participants

特征 Characteristics	总计 Total (n=25 871)	维生素D组 Vitamin D group (n=12 927)	安慰剂组 Placebo group (n=12 944)
女性 Female	13 085 (50.6)	6 547 (50.6)	6 538 (50.5)
年龄[a]/岁 Age/years	67.1±7.1	67.1±7.0	67.1±7.1
种族或民族 Race or ethnic group			
非西班牙裔白人 Non-Hispanic White	18 046/25 304 (71.3)	9 013/12 647 (71.3)	9 033/12 657 (71.4)
黑人 Black	5 106/25 304 (20.2)	2 553/12 647 (20.2)	2 553/12 657 (20.2)
非西班牙裔黑人 Non-Black Hispanic	1 013/25 304 (4.0)	516/12 647 (4.1)	497/12 657 (3.9)
亚洲人或太平洋岛民 Asian or Pacific Islander	388/25 304 (1.5)	188/12 647 (1.5)	200/12 657 (1.6)
美国印第安人或阿拉斯加原住民 American Indian or Alaskan Native	228/25 304 (0.9)	118/12 647 (0.9)	110/12 657 (0.9)
其他或未知 Other or unknown	523/25 304 (2.1)	259/12 647 (2.0)	264/12 657 (2.1)
BMI[a]/(kg·m^{-2})	28.1±5.7	28.1±5.7	28.1±5.8
糖尿病 Diabetes	3 537/25 824 (13.7)	1 804/12 900 (14.0)	1 733/12 924 (13.4)
目前吸烟 Current smoker	1 835/25 488 (7.2)	921/12 732 (7.2)	914/12 756 (7.2)
25-羟基维生素D水平[a] Level of 25-hydroxyvitamin D/(ng·mL^{-1})	30.7±10.0	30.7±10.0	30.7±10.0

【注】括号外为人数,括号内为百分比/%;a:$\bar{x}\pm s$。

[Note] Outside the brackets are the number of cases, and inside the brackets are proportions/%; a: $\bar{x}\pm s$.

在医学研究的统计报告中,除了用统计指标和统计表对变量进行综合描述,还可以用统计图直观地描述。例如,将连续变量绘制成散点图、直方图、箱线图或小提琴图,直观地评估正态性和潜在的离群值。临床上常出现的问题是用三维图形表示二维数据,虽然增加了美观,却违反了数据表达原则。

3 参数检验与非参数检验

在对医学数据中的变量进行描述性统计后,就开始进行正式的统计比较和检验假设。这时,须比较两组或多组之间处理因素或分组因素间某指标是否有差异以及这些差异是否为偶然误差[16]。在医学文献中,通常认为P<0.05是有统计学意义的。

对于比较组间差异的目标,检验方法的选择取决于所要比较数据的类型和分布情况。对于满足正态分布的连续数据使用参数检验,如t检验或ANOVA。当不满足正态分布以及数据为分类或序数数据时,可以使用非参数检验,如Wilcoxon秩和检验、χ^2检验等[17]。如果变量转换后的数据符合正态分布,那么对转换后的数据进行参数检验将比非参数检验更有效、更有意义[13]。例如,当抗体几何平均滴度遵循正态分布时,比较6个年龄组间的抗体几何平均滴度,则采用ANOVA;当比较不同年龄组间的抗体阳性率差别,则使用χ^2检验[18]。两组及多组比较的方法见表2。

正确的数据格式是分析的基础。根据不同的分析方法,将数据整理成统计软件可以识别分析的数据格式。t检验、ANOVA、Wilcoxon秩和检验(Mann-

表2　两组及多组比较的方法

Table 2　Methods of comparison between two and more groups

数据类型 Data type	两组 Two groups	两组以上 Two or more groups	配对 Pairs groups
定量资料 Quantitative data			
正态分布 Normal distribution	独立样本t检验 Independent samples t test	ANOVA	配对样本t检验 Paired-samples t test
非正态分布 Non-normal distribution	Wilcoxon秩和检验 Wilcoxon rank-sum test	Kruskall-Wallis秩和检验 Kruskall-Wallis test	Wilcoxon符号秩和检验 Wilcoxon signed-rank test
定性资料 Qualitative data			
分类资料 Categorical data	χ^2检验或Fisher's精确检验 χ^2 test or Fisher's exact test	χ^2检验 χ^2 test	McNemar's检验 McNemar's test
等级资料 Ordinal data	Wilcoxon秩和检验 Wilcoxon rank-sum test	Kruskall-Wallis秩和检验 Kruskall-Wallis test	Wilcoxon符号秩和检验 Wilcoxon signed-rank test

Whitney U)和Kruskal-Wallis秩和检验都是研究不同组间的差异,所以数据格式中必须包含组别信息和研究因素的测量值。配对数据是医学研究中常见的一种数据类型,其格式也相对较为特殊。常用统计分析方法的数据格式见图3。

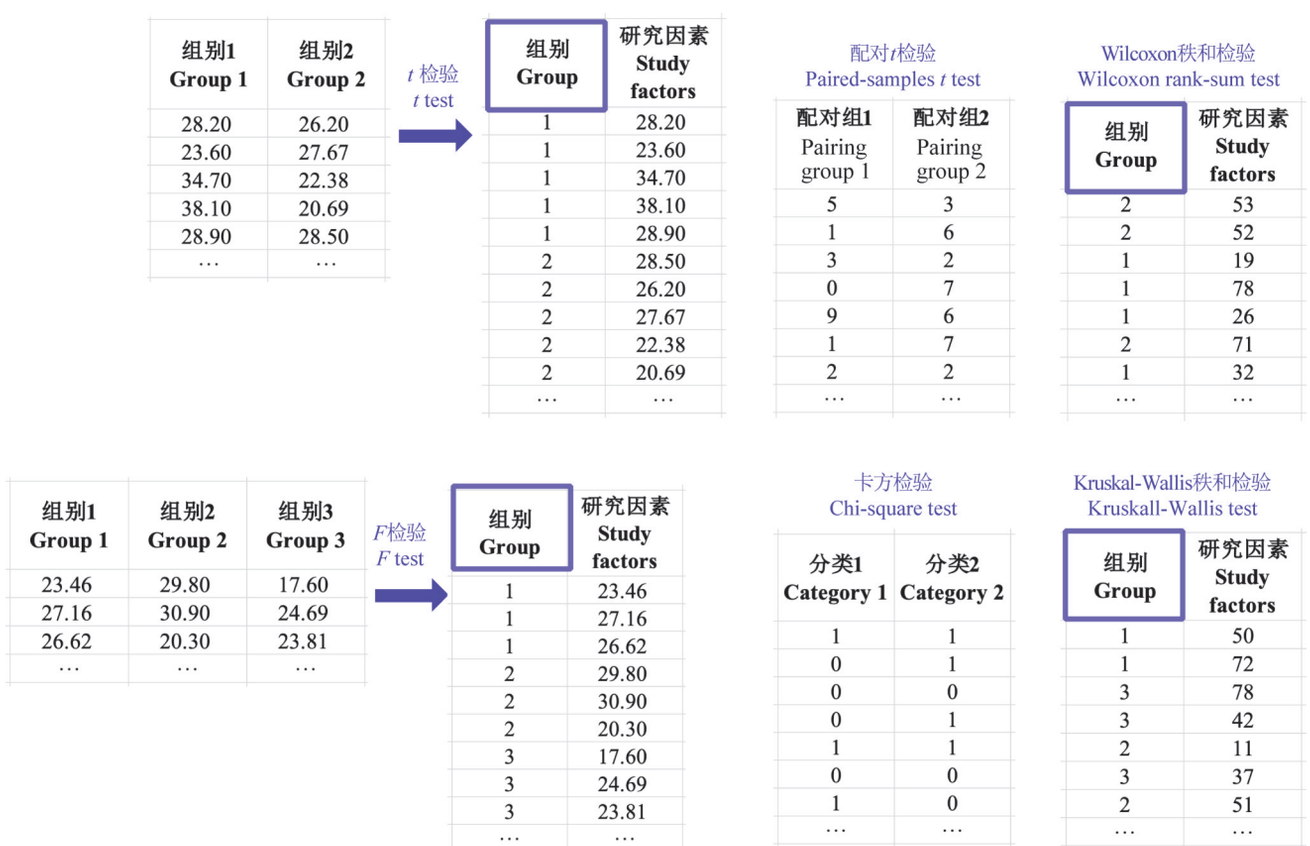

图3　常用统计分析方法的数据格式

Figure 3　Data format of common statistical analysis methods

由于一些检验假设的要求比较严格,包括数据的正态性和样本量,因此无法在所有情况下使用参数检验。在样本量较小的情况下,正态性检验方法对非正态性不太敏感,尽管有非正态性数据,但仍有机会检测到正态性。因此,对于小样本的数据,即使遵循正态分布,也建议使用非参数检验[19]。此外,参数检验方法还适用于大样本量的非正态分布数据,因为根据中心极限定理,随着样本量的增加,样本均值的分布趋于正态分布[20]。

4　相关性分析

医学研究的结果经常产生相互关联的数据(如吸

烟与肺癌)。相关分析与回归不同的是,它不用于预测,也无法验证因果关系,而是识别和衡量2个变量之间的线性关系的强度和方向。如果双变量满足近似正态分布,使用Pearson相关分析,否则应使用Spearman秩相关分析或Kendall秩相关分析进行评估。相关性用相关系数r表示,值范围从-1到1,概括了相关的强度和方向(0表示没有相关性,r值接近-1或1分别表示强的负相关性和正相关性)。在描述相关分析结果时,建议报告相关系数及其95%*CI*和*P*值。$P<0.05$代表2个变量之间存在相关关系。

例如,由于上海市闵行区手足口病周发病数与同期各气象因素(周平均温度、周平均相对湿度、周降水量等因素)满足近似正态分布,选择Pearson相关分析。将单因素相关分析中有统计学意义的气象因素纳入多重线性逐步回归分析中,研究气象因素变化与手足口病发生的关系[21]。

5 回归分析

回归分析在医学研究中应用非常广泛,包含4个主要方面:① 解释因变量(也称为响应变量、结果变量)和自变量(也称为协变量、预测变量和解释变量)之间的关系;② 识别、预测风险因素;③ 校正混杂因素;④ 建立预测模型[22]。在不允许随机分配到治疗组的情况下,多因素回归分析尤为重要,如在观察性研究中通过考虑潜在的混杂因素来确定暴露和结果之间的"独立"联系。除此之外,在前瞻性研究设计中还可用于自变量和因变量之间因果关系的推断[23]。

回归模型有不同的类型,这些模型的选择或使用应以研究问题和现有的数据为指导。医学研究中常用的回归分析方法包括线性回归、logistic回归、Cox回归和Poisson回归,分别采用连续数据、分类数据、生存数据和计数数据类型的结局指标。见表3。

表3 医学研究中常用的回归分析方法

Table 3 Regression analysis methods commonly used in medical research

数据类型 Data type	方法 Methods	效果估计 Effect estimation	假设条件 Assumptions
连续资料 Continuous data	线性回归 Linear regression	回归系数 Regression coefficients	线性、独立、正态和等方差 Linear, independent, normal and equal variance
分类资料 Categorical data	Logistic回归 Logistic regression	比值比 Odds ratio (*OR*)	观察值的独立性;连续型的自变量与因变量的对数值之间存在线性关系;多变量模型中自变量之间不存在多重共线性 Independence of observations; linear relationship between logarithmic values of independent and dependent variables in continuous type; no multicollinearity between independent variables in multivariate models
生存资料 Survival information	Cox比例风险回归 Cox proportional risk regression	风险比 Hazard ratio (*HR*)	比例风险(PH)不变 Proportional hazards (PH) constant
计数资料 Count data	Poisson回归 Poisson regression	风险比 Hazard ratio (*HR*)	事件独立发生;泊松分布 Events occur independently; Poisson distribution

在医学领域的许多实际问题中,结果受到不止一个因素的影响,如多种因素共同作用引起的疾病。在这种情况下,多因素分析是正确的选择。多因素分析方法包括多元线性回归分析、logistic回归及Cox回归模型(图4)。自变量的纳入,需要根据单因素分析、临床专业知识和现有文献来确定,即使其可能在统计学上没有意义[24]。共线性会影响回归模型拟合,需要对自变量进行共线性判断,评估指标有方差膨胀因子(variance inflation factor, VIF)、条件指数、Pearson相关系数和容忍度等指标。

5.1 线性回归模型

线性回归[25]是最常用的一种回归分析方法,假设一个连续型因变量和一个或多个自变量之间遵循线性关系,可分为简单线性回归和多元线性回归。多元线性回归有≥2个自变量,而简单线性回归通常只有1个自变量。多元线性回归模型是简单线性回归模型的扩展。例如一项研究[26]为估计韩国成年人与健康有关的体能(如手握力、肌肉耐力等),遵守多元线性回归模型的基本假设:线性、独立、等方差、连续性、正态性。其中,使用Kolmogorov-Smirnov检验来验证残差的分布情况。该研究采用了逐步回归的多元线性回归模型,回归系数来评估自变量对因变量的解释力。

5.2 Logistic回归模型

尽管线性回归模型十分常用,但它并不适用于某些类型的医学研究结果。对于二分类、多分类或有序的观察结果,如死亡或存活,通常选择logistic回归方法[27]。Logistic回归可以直接输出比值比(*OR*),用作估计效应大小的指标,以测量每个自变量与结局之间的关联强度[28]。此外,logistic回归也能用于估计因果推断的倾向得分。例如研究急诊手术时间(白天、晚上或

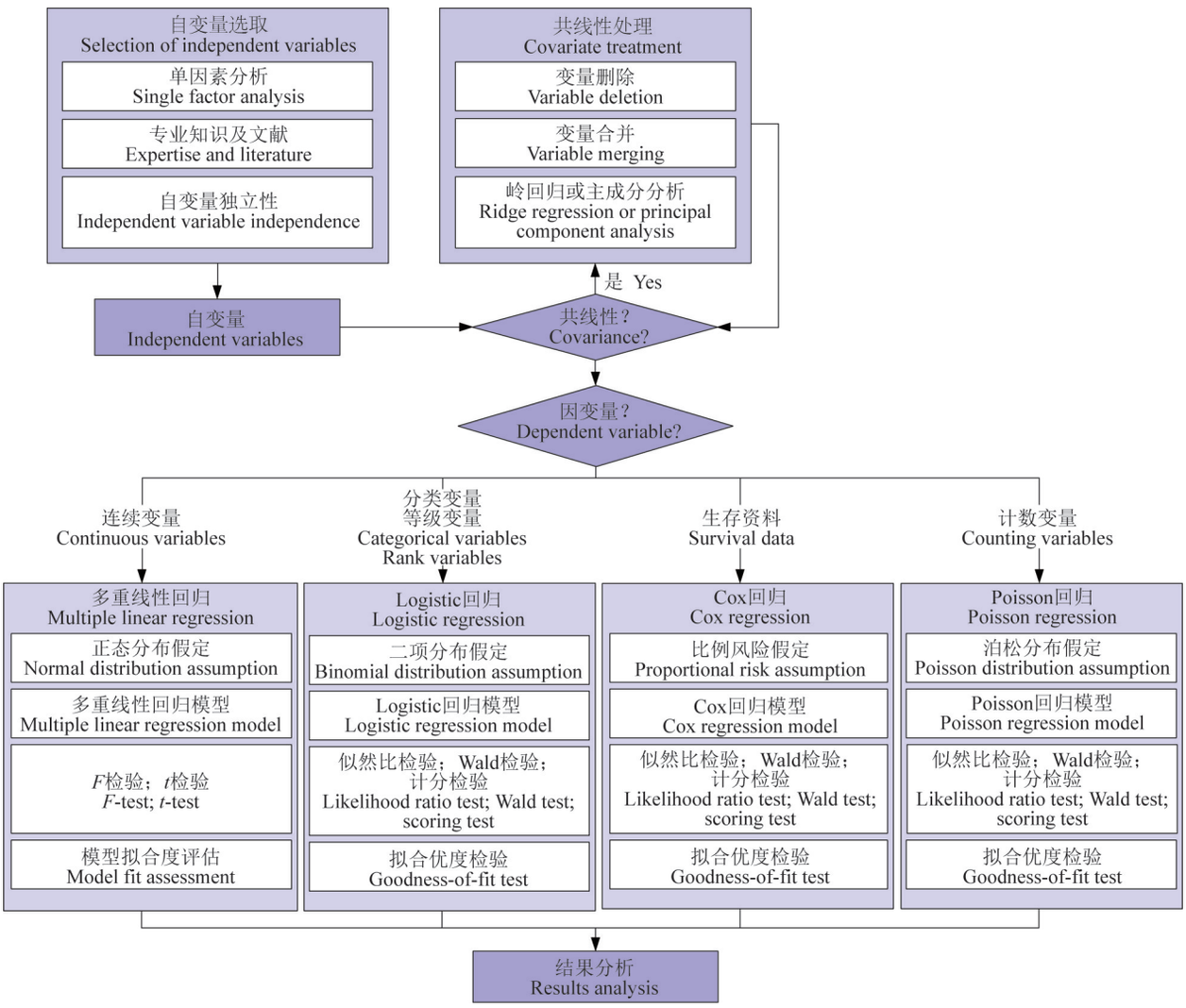

图4　多因素分析方法的选择

Figure 4　Selection of multi-factor analysis methods

夜间）与术后30 d住院死亡率的关系[29]，由于研究结局是二分类结局（是否死亡），采用二元logistic回归分析。考虑的协变量有年龄、性别、紧急类别、麻醉持续时间、手术类型等，同时描述了各变量的纳入形式。负二倍对数似然比检验结果显示，纳入这些变量的模型整体是有统计学意义的。该文章还使用了Box-Tidwell检验评估连续型自变量与因变量的对数转换值之间的线性关系，用Hosmer-Lemeshow拟合度χ^2检验来评估模型的拟合优度。

5.3　Cox回归模型

Cox比例风险回归模型是最常用的生存回归模型。Cox比例风险回归模型[30]可以同时解释多个协变量或风险因素，而在前面提到的Kaplan-Meier法与寿命表法只能进行单个分组变量的生存分析。Honigberg等[31]研究过早自然绝经和手术绝经与心血管疾病风险之间的联系。研究结局是时间-事件资料，故该研究采用Cox比例风险回归模型分析。这项研究纳入相关的混杂因素作为协变量，同时评估了连续型协变量的正态性，并对C-反应蛋白进行了对数转换以达到正态性。此外，使用Schoenfeld残差检验评估Cox回归比例风险的假设。

Cox回归模型由于不依赖特定的生存分布特点，已经在应用上取得了很大的普及，在处理再入院率、发病率和死亡率等混合结果的指标时具有较高的实用性和灵活性。只有满足比例风险（proportional hazards，PH）假设，Cox回归模型的结果才有效[32]。PH的基本假设为风险函数与基线风险函数的比值为固定值，即不随时间的改变而改变。Cox回归模型是否满足比例风险假设，可以通过Kaplan-Meier生存曲线图、Schoenfeld残差图和Schoenfeld残差检验等方法来检验。如果不满足比例风险的假设，模型中的自变量可随着时间的推移而相互影响，导致有偏倚的估计，甚至得出与问题的实际解释相矛盾的结论。当风险比例的假定条件不成立时，可采用拟合一个分层Cox模型，其

基线危险可以在各层之间不同，或者拟合一个含时间依存协变量 Cox 回归模型（time-dependent Cox regression model）。

5.4 Poisson 回归模型

Poisson 回归模型在医学研究中有重要作用，用于分析符合泊松分布的事件与影响因素之间的关系，且适用于罕见结局事件的研究。例如研究小儿视神经脊髓炎谱系疾病复发次数的影响因素[33]，由于因变量是计数资料，即使用 Poisson 回归分析。Poisson 回归模型还可用于队列研究来分析时间-事件数据[34]。研究发病率的一种方法是拟合 Poisson 回归模型，假设病例总数服从泊松分布。Poisson 回归模型是 Cox 模型的重要替代方法，因为它是估计累积暴露直观有效的方法，允许风险依赖于多个时间尺度估计生存分布函数和 *HR*。

5.5 回归模型的结果表达

回归模型应根据研究目的报告不同的内容[24]。例如，对于旨在建立预测模型的分析，还必须解释用于反映模型拟合度的指标。这些指标包括决定系数 R^2、均方根误差（root mean square error，RMSE）、赤池信息准则值（Akaike information criterion，AIC）、ROC 曲线下面积和 C 指数（concordance index，C-index）等。然而，对于变量选择，无论选择哪种方法，都必须明确描述选择程序[35]。

在线性回归分析结果中，建议报告回归方程、标准化回归系数 β、t 值和 P 值，一个衡量模型“拟合度”的指标（简单回归的决定系数 r^2 或多重线性回归的多重决定系数 R^2）以及多重共线性严重程度的指标（方差膨胀因子、条件指数和 Pearson 相关系数）等[36]。同时，logistic 回归通常需要报告 *P* 值、*OR* 值、95%*CI* 以及拟合优度指标（如 Hosmer-Lemeshow、Pearson χ^2 拟合优度指标）。Cox 回归、Poisson 回归的报告形式与 logistic 回归相似，但 *OR* 由 *HR* 和相对危险度（*RR*）代替[37]。

6 生存分析

由于生存时间的分布一般不服从正态分布，且通常含有删失数据，所以生存分析有其独特的统计方法。生存分析是对时间-事件（time-to-event）数据的分析[38]。这种数据描述了从一个起始事件到一个感兴趣的终点事件所经历的时间。

生存分析[39]是一系列的分析过程，包括统计描述、组间比较和回归分析（图 5）。通常用 Kaplan-Meier 曲线法描述小样本或大样本未分组的生存资料，应用寿命表法描述频数形式的生存资料和大样本分组的生存资料，估计总体生存率、中位生存时间及生存曲线等[40]；组间比较通常是比较不同群体或不同治疗措施的生存率，研究不同群体间或不同治疗措施的生存过程是否存在差异，通常使用 Log-rank 检验和 Breslow 检验比较两组或多组的生存率或生存曲线是否相同；回归分析探讨影响生存时间的因素以及在多大程度上影响有关事件的风险。

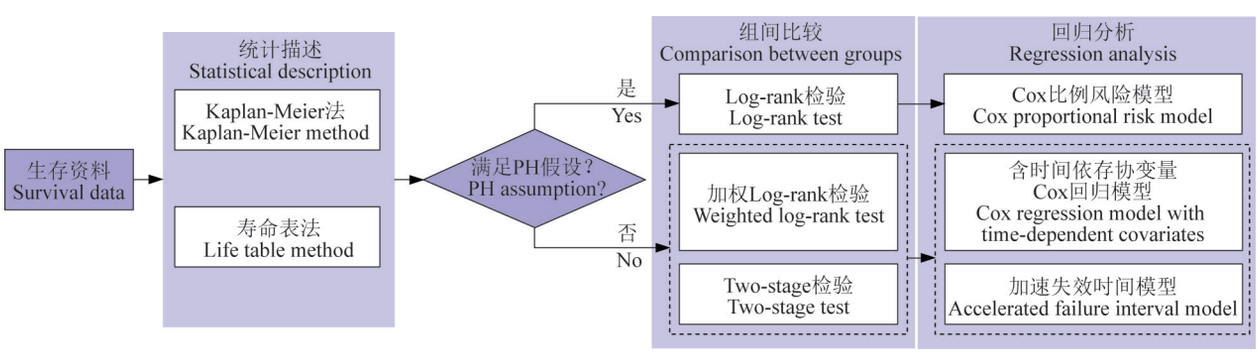

图 5 生存分析流程

Figure 5 Selection of multi-factor analysis methods

例如有研究将食管鳞癌患者分为膳食视黄醇低摄入组（N_1=62）、中摄入组（N_2=249）和高摄入组（N_3=77），3 组样本量均属于小样本量，故采用 Kaplan-Meier 曲线法估计生存率。因为 Kaplan-Meier 生存曲线无明显交叉出现，满足 PH 假设，所以应用 Log-rank 检验进行不同视黄醇摄入组的生存率的差异比较，采用 Cox 比例风险模型分析视黄醇和食管鳞癌患者预后的关联[41]。

Log-rank 检验对观察晚期的差异敏感，而 Breslow 检验对观察早期的差异敏感[42]。例如，利用 Breslow 检验和 Log-rank 检验对艾滋病感染者/艾滋病患者的近期和远期生存率进行统计学检验，分析短期和长期生存率差异是否有统计学意义[43]。当不满足 PH 时，比较组间生存率用 Log-rank 检验，其检验效能下降，应该考虑具有较好检验效能的加权 Log-rank 检验和 Two-stage 检验等方法[44]。

生存分析方法[45]一般分为参数法、半参数法和非

参数法3种类型。参数模型是在各种生存时间的分布概率模型基础上建立的，需要假设生存时间服从某种特定的分布，如Weibull分布、Gamma分布、指数分布和Log-logistic分布等。如果生存时间服从特定的参数分布，相比于非参数模型，参数模型能更好地拟合数据，显著提高统计效率并提供有意义的预测；半参数法模型最常用的是Cox比例风险模型；最简单和通用的方法是非参数方法，它不需要假设具体的分布形式，包括Kaplan-Meier方法和Breslow检验等。模型的选择应取决于是否符合模型的假设，如Cox比例风险回归模型和Log-rank检验的比例风险假设。加速失效时间模型（accelerated failure time model，AFT）是Cox回归模型的一个替代方案，可以为不满足比例风险假设的数据提供更好的估计方法，在存在删失数据的情况下，研究协变量与对数生存时间之间的线性回归关系，直观地解释分析结果[32]。与Cox回归模型不同，AFT模型通常是参数生存回归模型，如根据生存时间的Weibull分布拟合参数AFT生存回归模型。

7　统计符号的书写

统计学符号要按GB/T 3358.1—2009《统计学名词及符号第1部分：一般统计术语与用于概率的术语》的有关规定书写，以下符号均用斜体书写。样本的算术均数用英文小写$\bar{x}$，中位数用M；标准差用英文小写s；标准误用英文小写$s_{\bar{x}}$；t检验用英文小写t；F检验用英文大写F；卡方检验用希文小写χ^2；相关系数用英文小写r；自由度用希文小写ν；样本数用英文小写n；q检验用英文小写q；概率用英文大写P（P值前应给出具体检验值，如t值、F值等）；比值比用英文大写OR；95%置信区间用95%CI表示。

8　总结

医学数据的清洗及选择适当的统计方法对高质量的研究是非常重要的。本文对医学数据的清洗及常用统计方法的选择和应用进行了总结，旨在帮助医学研究人员熟悉更多的统计方法。统计方法的选择一般由研究目的和结果数据的类型决定。因此，建议医学研究人员在收集数据前咨询有经验的统计学工作者，制订一个全面的分析计划，并根据研究方案和统计分析的要求收集合格的数据资料。

（作者声明本文无实际或潜在的利益冲突）

参考文献

[1] MISHRA P, PANDEY C M, SINGH U, et al. Selection of appropriate statistical methods for data analysis[J]. Ann Card Anaesth, 2019, 22(3): 297-301.

[2] HE J, JIN Z C, YU D H. Statistical reporting in Chinese biomedical journals[J]. Lancet, 2009, 373(9681): 2091-2093.

[3] SEBASTIÃO Y V, PETER S DST. An overview of commonly used statistical methods in clinical research [J]. Semin Pediatr Surg, 2018, 27(6): 367-374.

[4] MATA D A, MILNER D A. Statistical methods in experimental pathology: a review and primer[J]. Am J Pathol, 2021, 191(5): 784-794.

[5] HAYAT M J, POWELL A, JOHNSON T, et al. Statistical methods used in the public health literature and implications for training of public health professionals[J]. PLoS One, 2017, 12(6): e0179032.

[6] 陈婉娇. 缺失数据插补方法及其在医学领域的应用研究[D]. 广州：华南理工大学，2019.
CHEN W J. Research on application of missing data imputation in medical field[D]. Guangzhou: South China University of Technology, 2019.

[7] 朱荣慧，许金芳，王睿，等. 多重填补技术在医学研究缺失值处理中的应用及发展[J]. 中国卫生统计，2022，39(2)：293-295，298.
ZHU R H, XU J F, WANG R, et al. Application and development of multi-fill technique in the processing of missing values in medical research[J]. Chin J Health Stat, 2022, 39(2): 293-295, 298.

[8] 花琳琳，施念，杨永利，等. 不同缺失值处理方法对随机缺失数据处理效果的比较[J]. 郑州大学学报（医学版），2012，47(3)：315-318.
HUA L L, SHI N, YANG Y L, et al. Comparison of different methods in dealing with missing values of missing at random[J]. J Zhengzhou Univ (Med Sci), 2012, 47(3): 315-318.

[9] 宋亮，万建洲. 缺失数据插补方法的比较研究[J]. 统计与决策，2020，36(18)：10-14.
SONG L, WAN J Z. Comparative research on interpolation method of missing data[J]. Stat Decis, 2020, 36(18): 10-14.

[10] 王晓晓，陶立元，裴敏玥，等. 连续变量转换为分类变量的几种方法[J]. 中华儿科杂志，2022，60(5)：420.
WANG X X, TAO L Y, PEI M Y, et al. Several methods for converting continuous variables into categorical variables[J]. Chin J Pediatr, 2022, 60(5): 420.

[11] KESTIN I. Statistics in medicine[J]. Anaesth Intensive Care Med, 2018, 19(3): 136-143.

[12] MISHRA P, PANDEY C M, SINGH U, et al. Descriptive statistics and normality tests for statistical data[J]. Ann Card Anaesth, 2019, 22(1): 67-72.

[13] FENG C Y, WANG H Y, LU N J, et al. Log transformation: application and interpretation in biomedical research[J]. Stat Med, 2013, 32(2): 230-239.

[14] VON ELM E, ALTMAN D G, EGGER M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) Statement: guidelines for reporting observational studies[J]. Int J Surg, 2014, 12(12): 1495-1499.

[15] LEBOFF M S, CHOU S H, RATLIFF K A, et al. Supplemental vitamin D and incident fractures in midlife and older adults[J]. N Engl J Med, 2022, 387(4): 299-309.

[16] BENSKEN W P, HO V P, PIERACCI F M. Basic introduction to statistics in medicine, part 2: comparing data[J]. Surg Infect, 2021, 22(6): 597-603.

[17] MISHRA P, PANDEY C, SINGH U, et al. Scales of measurement and presentation of statistical data[J]. Ann Card Anaesth, 2018, 21(4): 419-422.

[18] 马小敏，王芳，李淑华，等. 上海市金山区健康体检人群麻疹抗体水平监测结果分析[J]. 上海预防医学，2020，32(4)：320-324，328.
MA X M, WANG F, LI S H, et al. Surveillance of the levels of measles antibody in physical examination population in Jinshan District, Shanghai[J]. Shanghai J Prev Med, 2020, 32(4): 320-324, 328.

[19] GHASEMI A, ZAHEDIASL S. Normality tests for statistical analysis:

a guide for non-statisticians [J]. International Journal of Endocrinology and Metabolism, 2012, 10(2): 486-489.
[20] VRBIN C M. Parametric or nonparametric statistical tests: considerations when choosing the most appropriate option for your data [J]. Cytopathology, 2022, 33(6): 663-667.
[21] 赵霞,崔劲松,张奕,等.2014—2018年上海市闵行区手足口病流行特征及其与气象因子相关分析[J].上海预防医学,2022,34(3):219-222,234.
ZHAO X, CUI J S, ZHANG Y, et al. Epidemiological characteristics of hand, foot and mouth disease and its correlation with meteorological factors in Minhang District, Shanghai, 2014-2018[J]. Shanghai J Prev Med, 2022, 34(3): 219-222, 234.
[22] BENDER R. Introduction to the use of regression models in epidemiology [M]//VERMA M. Cancer epidemiology. Totowa, NJ: Humana Press, 2009: 179-195.
[23] 黄桥,黄笛,靳英辉,等.临床研究中常用的统计方法和常见问题[J].中国循证心血管医学杂志,2017,9(11):1288-1293.
HUANG Q, HUANG D, JIN Y H, et al. The common used statistic methods and common problems in clinical research[J]. Chin J Evid-Based Cardiovasc Med, 2017, 9(11): 1288-1293.
[24] SINGH S K, KAPLAN B, KIM S J. Multivariable regression models in clinical transplant research: principles and pitfalls [J]. Transplantation, 2015, 99(12): 2451-2457.
[25] SPERANDEI S. Understanding logistic regression analysis [J]. Biochem Med, 2014, 24(1): 12-18.
[26] KIM S W, PARK H Y, JUNG H, et al. Estimation of health-related physical fitness using multiple linear regression in Korean adults: national fitness award 2015-2019 [J]. Front Physiol, 2021, 12: 668055.
[27] TOLLES J, MEURER W J. Logistic regression: relating patient characteristics to outcomes[J]. JAMA, 2016, 316(5): 533-534.
[28] NORTON E C, DOWD B E, MACIEJEWSKI M L. Odds ratios-current best practice and use[J]. JAMA, 2018, 320(1): 84-85.
[29] TESSLER M J, CHARLAND L, WANG N N, et al. The association of time of emergency surgery-day, evening or night-with postoperative 30-day hospital mortality [J]. Anaesthesia, 2018, 73(11): 1368-1371.
[30] KARTSONAKI C. Survival analysis[J]. Diagn Histopathol, 2016, 22(7): 263-270.
[31] HONIGBERG M C, ZEKAVAT S M, ARAGAM K, et al. Association of premature natural and surgical menopause with incident cardiovascular disease[J]. JAMA, 2019, 322(24): 2411-2421.
[32] GEORGE B, SEALS S, ABAN I. Survival analysis and regression models[J]. J Nucl Cardiol, 2014, 21(4): 686-694.
[33] ZHANG S C, QIAO S, LI H Y, et al. Risk factors and nomogram for predicting relapse risk in pediatric neuromyelitis optica spectrum disorders[J]. Front Immunol, 2022, 13: 765839.
[34] AGRESTI A. An introduction to categorical data analysis[M]. 3rd ed. Hoboken: John Wiley & Sons, 2018.
[35] ALTHOUSE A D, BELOW J E, CLAGGETT B L, et al. Recommendations for statistical reporting in cardiovascular medicine: a special report from the American heart association[J]. Circulation, 2021, 144(4): e70-e91.
[36] LANG T A, ALTMAN D G. Basic statistical reporting for articles published in biomedical journals: the "statistical analyses and methods in the published literature" or the SAMPL guidelines[J]. Int J Nurs Stud, 2015, 52(1): 5-9.
[37] FENG G S, QIN G Y, ZHANG T, et al. Common statistical methods and reporting of results in medical research [J]. Cardiovasc Innov Appl, 2022, 6(3): 117-125.
[38] TOLLES J, LEWIS R J. Time-to-event analysis[J]. JAMA, 2016, 315(10): 1046-1047.
[39] BENÍTEZ-PAREJO N, RODRÍGUEZ DEL ÁGUILA M M, PÉREZ-VICENTE S. Survival analysis and Cox regression [J]. Allergol Immunopathol, 2011, 39(6): 362-373.
[40] BARAKAT A, MITTAL A, RICKETTS D, et al. Understanding survival analysis: actuarial life tables and the Kaplan-Meier plot[J]. Br J Hosp Med, 2019, 80(11): 642-646.
[41] 曾巧燕,张菊薇,王见文,等.膳食视黄醇与食管鳞癌患者预后的关联研究[J].上海预防医学,2022,34(11):1085-1089.
ZENG Q Y, ZHANG J W, WANG J W, et al. Association of dietary retinol with prognosis in patients with esophageal squamous carcinoma [J]. Shanghai J Prev Med, 2022, 34(11): 1085-1089.
[42] MARTINEZ R L M C, NARANJO J D. A pretest for choosing between logrank and Wilcoxon tests in the two-sample problem[J]. Metron, 2010, 68(2): 111-125.
[43] 辅海平,孙强.浙江省桐乡市2005—2019年HIV/AIDS患者CD4$^+$T淋巴细胞检测情况分析[J].中国皮肤性病学杂志,2021,35(9):1007-1011.
FU H P, SUN Q. Analysis of CD4$^+$ T lymphocyte test in patients with HIV/AIDS in the Tongxiang City of Zhejiang Province from 2005 to 2019[J]. Chin J Derm Venereol, 2021, 35(9): 1007-1011.
[44] 刘宏,许燕波,何雪心.非比例风险模型生存分析方法的选择及应用[J].中国卫生统计,2020,37(1):127-131.
LIU H, XU Y B, HE X X. Selection and application of non-proportional risk model survival analysis methods[J]. Chin J Health Stat, 2020, 37(1): 127-131.
[45] 陈兵,骆福添.生存分析中的回归模型[J].中国卫生统计,2006,23(5):462-465.
CHEN B, LUO F T. Regression models in survival analysis[J]. Chin J Health Stat, 2006, 23(5): 462-465.

（**收稿日期**：2022-12-16；**网络首发**：2023-06-06）
（**中文编辑**：伦宜然；**英文编辑**：巩婧恬；**校对**：符移才）