# 医学论文中 logistic 回归分析的误用及案例分析*

韩宏志 官 鑫 李欣欣 姜瑾秋 王 丽†
《吉林大学学报(医学版)》编辑部,130021,长春

**摘 要** logistic 回归分析在医学研究领域应用较为广泛,但因其涉及的因素较多,统计方法较为复杂,部分研究者未咨询流行病学和卫生统计学专业人员而盲目套用统计学方法,造成其在应用 logistic 回归分析处理数据时出现诸如自变量选取和纳入、统计模型选择、结果表述和报告时误用错用情况的发生。本文选取公开发表的医学论文,举例分析文章中 logistic 回归误用错用的情况,指出其正确应用方法,为医学期刊编辑处理类似稿件提供参考。

**关键词** 医学论文; logistic 回归;危险因素;误用

**Misuse of logistic regression analysis and case analysis in medical papers**//HAN Hongzhi, GUAN Xin, LI Xinxin, JIANG Jinqiu, WANG Li

**Abstract** Logistic regression analysis is widely used in the field of medical research such as the screening of epidemic risk factors, etiological analysis, evaluation of clinical efficacy and the probability of predicting the incidence of disease. However, there are many factors involved in logistic regression analysis, and the statistical methods are complex. Some researchers are using statistical methods without consulting the epidemiologist and health statistics professionals, and thus result in the occurrence of misuse of logistic regression analysis, such as the selection and inclusion of independent variables, the selection of statistical models, and the result expression and report. We analyze the missuse of logistic regression analysis in the published papers and suggest the related correct formulations. We hope to provide the references for the medical journal editors to process the similar manuscripts.

**Keywords** medical paper; logistic regression analysis; risk factor; misuse

**Authors' address** Editorial Board of Journal of Jilin University (Medicine Edition), 130021, Changchun, China

**DOI**:10.16811/j.cnki.1001-4314.2018.06.009

logistic 回归分析可处理医学研究中反应变量为二项分类、多项有序分类和多项无序分类变量时反应变量与诸多自变量间相互关系的问题,是一种较为复杂的高级统计学应用方法。但由于研究者统计学基础知识不牢固,在整体实验设计和应用 logistic 回归分析进行数据分析时对诸多因素处理不当而往往盲目套用统计学方法,在模型假设判断、样本量估计、变量赋值、回归分析结果表达和描述、自变量筛选方法选择及检验水准设定等方面存在诸多问题。本文结合一些实例进行论述,为论文作者、审稿专家、医学期刊编辑正确应用 logistic 回归分析方法处理相关数据提供参考和借鉴。

## 1 回归分析中赋值存在的问题

**例 1** 取自文献[1]。原文回顾性分析行双侧脑脑硬脑膜颞浅动脉血管融通术 EDAS 治疗、术后采用 DSA 随访的出血型烟雾病患者的临床资料,分析项目包括性别、首次手术年龄、术前出血次数、脑代谢情况评估、脑出血类型、是否合并脑缺血症状、烟雾病分期、大脑后动脉是否受累及是否有并发症(高血压、高血脂、高血糖任意一项),探讨这些因素对血管重建效果的影响。采用单因素分析和多因素 logistic 回归分析可能影响出血型烟雾病患者术后血管重建的因素,并对纳入 logistic 回归分析的主要研究因素进行赋值,见表 1[1]。

**表 1 主要研究因素和赋值方法**

| 变量 | 赋值方法 |
|---|---|
| 年龄 | <18 岁=0,18~30 岁=1,31~40 岁=2,41~50 岁=3,>50 岁=4 |
| 出血次数 | 未出血半球=0,出血 1 次=1,出血 2 次=2,≥3 次=3 |
| 是否合并缺血 | 不合并=0,合并=1 |
| 脑代谢情况 | 正常=0,代谢减低=1,代谢缺损=2 |

表 1 中为原文作者对自变量的赋值情况,其中年龄、出血次数和脑代谢情况为多分类变量,原文作者在进行赋值时采用了 0、1 和 2 等连续赋值,这样赋值一种自变量引入回归方程时仅能得到一个 OR 值,用来解释多分类变量的变化关系及其对因变量的影响,对结果的分析是有欠缺的。一般情况下,如年龄为连续性变量,赋值时可采用 10 岁为一个年龄段划分进行赋值,得到的回归系数就可以解释为年龄每增加 10 岁对因变量的影响,这种赋值方法是基于年龄与因变量间存在线性关系,并且该种疾病在各个年龄段间发病率相似,但实际情况下很难确定,应建议作者对这个年龄变量进行哑变量化。脑代谢情况中的 3 个分类(正常、代谢减低和代谢缺损)应属于无序多分类变量,不属于疾病严重程度的轻、中和重度,如进行 0、1 和 2 赋

* 吉林省卫生计生委 2017 年科技能力提升项目(2017G014,2017G015)
† 通信作者

值,无法合理有效地进行结果解释,也应该哑变量化。这样处理时,每个哑变量都能得出一个估计的回归系数,从而使得回归结果更易于解释,更具有实际意义。

变量的赋值不同对结果的解释也不同,只有正确进行自变量的赋值,才能正确解释暴露因素与研究结局之间的关系。作者在进行哑变量量化时还应考虑样本量大小,如果样本量足够大,可以将哑变量量化后引入回归模型,这样可以得到不同级别的差异,更直观地反映出该自变量的不同属性对于因变量的影响,提高了模型的精度和准确度;如果样本量不够大,过度赋值哑变量会造成变量数目增多,使得回归分析的结果变得不可靠。logistic 回归分析因变量赋值时,其中二分类变量的因变量赋值为 0 和 1(0 为阴性结果,1 为阳性结果),有序多分类变量自变量赋值可按照等级秩次赋值或哑变量赋值,无序多分类变量自变量赋值应进行哑变量赋值。连续型变量自变量赋值应根据研究目的和专业知识综合考虑是否进行向分类变量的转化,而有些研究者为了研究方便将连续型变量量化成分类变量引入回归方程,可能会损失数据中包含的有效信息,降低检验效率,外审专家和编辑应着重注意这方面的问题。

## 2　回归分析中样本量存在的问题

**例 2**　取自文献[2]。采用单因素分析和多因素 logistic 回归分析影响肠杆菌科细菌血流感染预后的相关影响因素。见表 2 和 3。

**表 2　肠杆菌科 BSI 患者预后的单因素分析[$n$(%)]**

| 存活影响因素 | 肠杆菌科死亡组 ($n$=15) | 肠杆菌科存活组 ($n$=20) | $\chi^2/t$ | $P$ |
|---|---|---|---|---|
| 年龄($\bar{x}\pm s$,岁) | 76.1±10.2 | 58.3±9.4 | 8.65 | <0.01 |
| 血清白蛋白 <3.5 g/L | 11(73.3) | 7(35.0) | 5.04 | <0.05 |
| APACHE Ⅱ ($\bar{x}\pm s$,分) | 23.2±5.3 | 15.4±4.2 | 7.95 | <0.01 |
| 休克 | 8(53.3) | 4(20.0) | 4.23 | <0.05 |
| MOSF | 7(46.7) | 3(15.0) | 4.21 | <0.05 |
| 中心静脉置管 | 12(80.0) | 9(45.0) | 4.38 | <0.05 |
| 气管插管或切开 | 11(73.3) | 6(0.3) | 6.44 | <0.05 |
| ESBLs 阳性检出率 | 13(86.7) | 8(40.0) | 7.78 | <0.01 |

**表 3　影响肠杆菌科 BSI 患者预后 logistic 回归分析**

| 存活影响因素 | logistic 回归分析 | | | | |
|---|---|---|---|---|---|
| | S.E | Wald$\chi^2$ 值 | $P$ | OR 值 | 95% CI |
| 年龄 | 0.534 | 4.979 | 0.043 | 5.442 | 1.977~7.783 |
| APACHE Ⅱ | 0.612 | 5.785 | 0.031 | 7.532 | 2.147~13.376 |
| 气管插管或切开 | 0.633 | 5.247 | 0.037 | 6.224 | 1.875~6.379 |
| ESBLs 阳性检出率 | 0.751 | 6.833 | 0.021 | 8.661 | 2.634~19.247 |

logistic 回归分析的统计推断建立在足够样本量基础上,为了得到可靠的参数估计,需要足够的样本量来保证参数估计的稳定性;经验估计样本量为协变量个数的 10~15 倍,本次研究单因素共有 8 个自变量,为满足 logistic 回归分析统计需要,病例组人数应为 80~120 例。原文中肠杆菌科死亡组和存活组总计 35 例,样本量过小,原文中肠杆菌科死亡组和存活组中血清白蛋白、休克、合并多器官功能不全、中心静脉置管、气管插管或切开和 ESBLs 阳性检出率的结局变量的个数均少于 10 个,由此计算出的单因素分析及 logistic 回归分析的结果和所得出的结论可信度下降,具有很大偏差,不具有说服力,无发表价值。

## 3　回归分析时统计模型选择存在的问题

**例 3**　取自文献[3]。原文探讨冠心病的危险因素,将冠心病作为因变量,BMI、合并高血压病、糖尿病、冠心病家族史、吸烟年支、血清 TC、TG、HDL-C、LDL-C 以及 Hcy 等作为自变量,进行多元 logistic 逐步回归后退法分析。

原文中因变量冠心病为二分类反应变量,即是否患有冠心病为一个因变量,多重 logistic 回归指仅有一个因变量,有 $k(\geq 2)$ 个自变量,而多元 logistic 回归指同时考察的因变量的个数≥2,因此原文中多元 logistic 回归的统计学分析方法说法是错误的,统计学方法应为二分类反应变量的 logistic 回归分析。原文采用筛选自变量的方法为逐步回归后退法的说法是错误的,一般 logistic 回归分析筛选自变量的方法主要有向前选择法、向后剔除法、逐步筛选法和最优子集法,对于采用逐步法筛选变量,还应该写明纳入标准 $\alpha_{入}$ 和剔除标准 $\alpha_{出}$,保留在模型中的变量的 $P$ 值应小于剔除变量的检验水准,原文中作者未提及 $\alpha_{入}$ 和 $\alpha_{出}$。选用不同的变量筛选方法得到的模型可能会不同,变量筛选时首先要进行专业上的考虑,重要的自变量不能遗漏,有些情况下可以多采取几种方法并结合专业知识进行判定,专业上无关的自变量也不能强行引入分析。

## 4　回归分析模型拟合优度检验存在的问题

当通过 logistic 回归分析建立初步模型后,并不代表分析的结束,而是需要对模型的合理性从统计学和专业角度进行评价。对所拟合的模型进行评价,即评价模型的预测值是否与观测值具有较高的一致性,logistic 回归模型的拟合优度是通过比较模型预测结果与实际观测时间发生与不发生的频数有无差别来进行检验的。模型的假设检验只能说明建立的模型是否具有统计学意义,不能描述模型的拟合效果。预测结果与实际观结果测越吻合,说明模型的拟合效果越好。一般采用偏差检验、Pearson $\chi^2$ 检验和 Homser-Leme-

show 统计量。在实际的情况中作者很少进行此种检测,编辑应注意作者是否进行模型评价。张英英等[4]检索收集了2010—2014年5种中华系列杂志发表的281篇多因素 logistic 回归文献,其中仅有28篇(9.96%)进行了模型效果评价。编辑应提醒并建议作者根据因变量和自变量的数据类型、专业上的意义及研究目的重新进行此种检验,以得到最优化的模型。

## 5 回归分析结果表达和描述存在的问题

logistic 回归分析结果描述时,检验统计量的值应给出回归系数的最大似然估计值($b$)、标准化回归系数的最大似然估计值($b'$)、Wald 检验的统计量值(Wald $\chi^2$)、Wald 检验统计量的自由度($df$)、Wald $\chi^2$ 检验的 $P$ 值、优势比 OR 的估计值(OR)及优势比 OR 的95%置信区间的下限和上限(OR 95% CI)。孙宇姣等[5]撰写的《应用 CT 评估冠心病危险因素与冠状动脉病变性质的关系》结果中对 logistic 回归分析结果的表述中仅报道了 OR(95% CI)和 $P$ 值,而龚丽英等[6]撰写的《多个危险因素对冠状动脉粥样硬化性心脏病的预测价值》结果中仅报道了 $b$ 值和 $P$ 值,缺乏对 logistic 回归分析结果的有效表述。不仅要给出 OR 值、$P$ 值,还应给出95% CI、检验方法、自变量筛选方法和检验统计量的值,以便对结果进行专业上的解释。给出相应的回归系数,可将多重 logistic 回归方程列出,以此明确表示各个自变量对结果变量的影响。

## 6 回归分析自变量的选取及纳入存在的问题

一般情况下,在 logistic 回归分析中自变量选择时应先进行单因素分析,找出有统计学意义的指标再进行 logistic 回归分析;但在实际应用中,由于自变量间可能存在一定程度的交互作用,即每个自变量对结果变量的影响贡献受到其他自变量的影响,而单因素分析时未考虑各因素间可能存在的交互作用,在单因素分析中存在统计学意义的自变量纳入回归方程时可能会无统计学意义,而单因素分析时无统计学意义的自变量引入 logistic 回归分析并非没有意义,单因素分析时使用传统的检验水准($P<0.05$)可能会掩盖一些重要的自变量,单因素分析的结果在很多情况下是不可靠的,其结果只是起参考作用,并不一定反映其真实效应。因此一般情况下当自变量个数较少可以直接将自变量不进行单因素分析,或根据前期研究将可能存在有生物学联系的变量直接带入回归方程;自变量个数较多时,可以放宽检验水准对变量进行筛选,选择单因素分析中显著性检验水平($P$)≤0.25 的变量纳入回归方程[7]。自变量的筛选即要考虑单因素分析中是否有统计学意义,还要从专业上考虑,对于专业上有意义的变量,即使单因素分析无统计学意义,也可以进入多因素分析的回归方程中,或者采用多种变量筛选技术,同时考虑因素之间的交互作用进行综合分析,得到较为可靠的结果。张英英等[4]的研究显示:3.91%(11/281)的多因素logistic回归文献进行了自变量间交互作用的统计分析,仅有1.78%(5/281)的多因素 logistic 回归文献进行了自变量间交互作用的分析,其他文章作者仅依据经验和专业知识等进行判断。编辑在审阅 logistic 回归分析的稿件时应注意选取的自变量间是否存在交互作用或相关性,应用统计学方法进行检验时应对自变量进行有效取舍,以免引入的自变量对统计结果产生偏差。

## 7 结束语

logistic 回归分析是医学科学研究中应用广泛且较为复杂的统计学方法,无论在专业角度还是在医学统计学应用方面都具有专业性,应用过程中易出现错用误用情况。编辑应建议作者在进行模型假设判断、自变量选取及赋值、样本量估计、变量纳入和统计学方法选择、检验水准设定、结果表述和报告等时咨询流行病学和卫生统计学专业人员,不要简单盲目地套用统计学方法,应辨别统计学方法的适用条件,结合专业知识确定结论,从而得出科学准确的结果和结论。

## 8 参考文献

[1] 刘鹏,李德生,杨伟中,等. 脑硬脑膜颞浅动脉血管融通术治疗出血型烟雾病的疗效及其影响因素分析[J]. 中国脑血管杂志,2013,10(4):169

[2] 孙伏喜,冯旰珠,高天明,等. 综合重症监护病房肠杆菌科细菌致血流感染发生的危险因素[J]. 中华医学杂志,2014,94(9):684

[3] 徐群威,赵微燕. 血清同型半胱氨酸水平与老年冠心病的相关性研究[J]. 中国医师杂志,2014(4):516

[4] 张英英,周晓彬,张健,等. 中文临床医学期刊中多因素 logistic 回归文献报告质量评价[J]. 中国公共卫生,2016,32(5):720

[5] 孙宇姣,俞鑫,耿松,等. 应用 CT 评估冠心病危险因素与冠状动脉病变性质的关系[J]. 中国医师杂志,2017,19(7):1069

[6] 龚丽英,彭丽萍,江凤林,等. 多个危险因素对冠状动脉粥样硬化性心脏病的预测价值[J]. 中国医师杂志,2013,15(9):1170

[7] 刘宏杰. Logistic 回归模型使用注意事项和结果表达[J]. 中国公共卫生,2001,17(5):466

(2018-07-04 收稿;2018-08-15 修回)