医学期刊编辑对多因素分析的核查要点

罗云梅 蒲素清 李缨来†

四川大学华西医院《中国普外基础与临床杂志》编辑部，610041，成都

摘 要 目前医学期刊中多因素分析方法的应用十分普遍。在处理此类稿件时，编辑需审查多因素分析方法的选用是否正确、变量条件是否满足、自变量纳入方法和选择是否合理、分析结果是否呈现完整、结果数据之间是否有矛盾等。文章详细阐述了多因素分析方法的关键审查要点，以供编辑同人参考。

关键词 医学期刊；多因素分析；审查要点

Editors of medical journals need to check key points of multivariate analysis//LUO Yunmei, PU Suqing, LI Yinglai

Abstract Multivariate analysis is common in the Chinese medical journal. While we deal with paper related this kind of method, we should pay attention on these points: whether the multivariate analysis is appropriate, whether multivariate analysis method meet the method's condition, whether the variable enrolled method is right and the selected variables are suitable, whether the analysis results is complete and data is right. This paper elaborates on these key points above, and we hope to make reference to the editors of Chinese medical journal.

Keywords medical journal; multivariate analysis; key point

Authors' address Editorial Department of Chinese Journal of Bases and Clinics in General Surgery, 610041, Chengdu, China

DOI:10.16811/j.cnki.1001-4314.2020.03.026

目前，医学期刊中统计方法采用多因素分析方法的论文越来越多，这是由于，多因素分析方法可同时分析多种因素的影响，有利于控制混杂因素的干扰[1]。目前医学期刊中的多因素分析方法有：多重线性回归分析、条件和非条件 logistic 回归、聚类分析、判别分析、主成分分析、Cox 比例风险回归模型、结构方程模型分析等。我们发现作者来稿中多因素分析内容往往存在或多或少的问题，如样本量过小、变量纳入方法有误、统计量和 P 值之间不对应等[2]；因此，有必要总结多因素分析方法的核查要点，为医学期刊编辑提供参考。本文仅总结最常见的多因素分析方法，包括多重线性回归分析、logistic回归及 Cox 比例风险回归模型，结合我们既往处理的稿件进行详细阐述。此外，本文的检验水准 α 默认为 0.05（双侧）。

1 多因素分析适用条件的核查

1.1 核查样本量是否足够

关于多因素分析方法包括多重线性回归分析、logistic回归及 Cox 比例风险回归模型的样本量估计问题，常用的经验估计方法是，样本量应是自变量数量的 5～20 倍甚至更多[3]；因此，编辑遇见此类分析时，首先即应核查样本量，样本量太小，几乎没有做多因素分析的必要，且此时多因素分析的模型拟合效果往往不理想。我们在编辑《MRI 检查预测乳腺癌新辅助化疗后病理完全缓解的准确性分析》一文时发现，作者在比较假阴性（21 例）和真阴性（12 例）患者的 MRI 参数时，进行了多因素 logistic 回归分析，纳入了多个变量，恐存在样本量不足的问题，提请作者核查，作者经过慎重考虑后删除了此部分的多因素分析结果。

1.2 核查是否满足多因素分析方法对于变量的要求

不同的多因素分析方法，对于变量都有其适用条件。多重线性回归的变量要求是，因变量是计量资料，自变量类型不限；logistic 回归模型的变量要求是，因变量是分类资料，包括二分类、无序多分类和有序多分类资料，而自变量的类型不限。

如文献[4]，作者欲探讨中央区淋巴结转移的影响因素，即采用了非条件 logistic 回归模型。Cox 比例风险回归模型适用于生存资料（包括生存时间和生存结局），而自变量的类型不限[1,3]。又如在文献[5]，研究指标——股动脉内膜剥脱术后再狭窄资料中，包括了再狭窄的发生情况和发生再狭窄的时间，是典型的生存资料，故该研究采用 Cox 比例风险回归模型进行多因素分析。

1.3 核查模型中纳入的变量是否合理及完整

1.3.1 *核查变量的纳入选择方法是否正确* 我们在编辑工作中发现，作者最常用的变量纳入方法为，根据单因素分析结果，选择有统计学意义的变量纳入多因素分析模型。我们认为，此种方法只正确纳入了部分的自变量，而合理的变量纳入方法是，根据单因素分析结果、专业知识及既往文献结果选择需纳入的变量。如在《结直肠癌患者术前肠道微生态失衡现状及影响因素分析》[6]一文中，作者除了选择单因素分析中 $P<0.1$ 的因素，包括婚姻状况、BMI 和平均每年使用抗生素时间，还根据专业考虑纳入了肠道准备和新辅助放化疗因素，进行了有序 logistic 回归。

目前大多认为，肿瘤的分化程度和预后密切相关，但若某研究者进行单因素分析时，发现肿瘤的分化程

† 通信作者

度没有统计学意义,据此认为多因素分析时不用纳入该变量,则是不合适的。单因素分析时没有统计学意义的原因在于,只考虑了肿瘤的分化程度这一个因素,其他因素未考虑在内。有可能在多因素分析时,在控制了其他混杂因素的影响后,肿瘤的分化程度对预后的影响才体现出来,才有统计学意义;因此,单因素分析结果不能作为多因素分析变量选择的唯一依据,且同研究的同一因素的单因素分析结果和多因素分析结果可能相悖。此外需强调的是对于配对资料进行条件 logistic 回归分析时,配对的变量不应纳入多因素分析。

1.3.2 *核查是否考虑了变量之间的相互影响*　以多重线性回归模型为例,在模型拟合过程中,需考虑变量之间的相互影响,即考虑变量之间的共线性问题,需进行共线性诊断[7]。共线性会对模型的拟合产生不良影响,如参数估计的精度降低、置信区间长度增宽[8]、模型拟合后系数解释困难等,应该避免。编辑在处理稿件过程中,若作者未提供相应的指标(如反应共线性严重程度的方差膨胀因子[9]、条件指数、方差分量等),可通过其他指标窥见一斑。如同时纳入了身高、体质量及体质量指数(BMI),或同时纳入了 T 分期、N 分期、M 分期及 TNM 分期时,务必重点考察各变量的解释是否符合专业结论。如出现了与专业解释相悖的情况(如回归系数本应为正值结果却为负值),或者某些指标的 Wald χ^2 值尤其大(超过 1 000 甚至几千),则需提请作者考察变量之间的共线性关系。最常见的处理共线性的方法为:1)从一组高度相关和具有多重共线性的自变量中删除某个变量[9],如 T、N、M、TNM 分期,可考虑只纳入 TNM 分期或者作者关注的那个因素,再建立回归模型;2)改变自变量的定义形式,将 2 个有多重共线性的自变量合并成一个变量或进行变量变换;3)进行岭回归或采用主成分分析。logistic 回归模型[10-11]和 Cox 比例风险模型[12-13]的共线性诊断和处理方法原理与多重线性回归类似,在此不再赘述。

2　多因素分析结果方面的核查

2.1　核查分类资料是否给出了对照

在多因素分析过程中,对于以分类资料形式纳入的变量,包括二分类资料和无序多分类资料,甚至是等级资料,都会设置一类对照。比如《结直肠癌同时性腹膜转移影响因素的多因素分析》[14]一文中,作者原文提供的非条件 logistic 回归结果,见表 1。以合并糖尿病为例(初投稿件中作者并未提供各分类变量的对照),因变量是关注的腹膜转移,对照不明的情况下,无法判定合并糖尿病患者的腹膜转移风险到底是高于还是低于未合并糖尿病患者。此外,补充一点,假设以合并糖尿病为对照,回归系数为 0.867,而若以未合并糖尿病为对照,回归系数则为 -0.867,符号刚好相反。若文章没有给出对照,也没有做任何文字说明,则对系数的解释就无从下手;因而,我们建议在编修过程中,要求作者提供变量赋值表,并在多因素分析表格内明确阐明对照类别。

我们认为,多因素分析时,呈现变量赋值表的作用在于:1)呈现各变量的赋值情况;2)呈现各变量的纳入形式,特别是等级资料;3)因变量为二分类变量时,呈现关注的结果(如 1 为患病时,关心患病;如 0 为患病、1 为未患病,则关心的是未患病),自变量为二分类变量时,呈现对照(程序分析时往往默认对照是赋值为 0 的特征)。如文献[14]中,作者呈现了因变量和自变量的赋值表(表 2),编辑对于因素的赋值便一目了然。

表 1　多因素非条件 logistic 回归分析结果

因　素	β 值	SE	Wald χ^2 值	OR 值	OR 95% CI	P 值
CEA 水平	1.694	0.626	4.883	5.442	1.169 ~ 13.592	<0.05
CA-125 水平	3.224	0.631	17.028	25.128	3.923 ~ 46.534	<0.05
T 分期	-2.338	0.890	15.223	0.096	0.005 ~ 0.178	<0.05
分化程度	-1.298	0.412	3.224	0.273	0.123 ~ 0.606	<0.05
病理学类型	-1.617	0.323 5	11.554 5	0.198	0.058 ~ 0.680	<0.05
合并糖尿病	0.867	0.683	4.588	2.379	1.132 ~ 16.462	<0.05

表 2　变量赋值表

变量	赋值
腹膜转移	腹膜转移 = 1,无腹膜转移 = 0
年龄	≥65 岁 = 1, <65 岁 = 2
肿瘤直径	≥5 cm = 1, <5 cm = 2
肿瘤位置	右半 = 1,左半 = 2,直肠 = 3
CEA 水平	正常 = 0,增高 = 1
CA19-9水平	正常 = 0,增高 = 1
CA-125 水平	正常 = 0,增高 = 1
T 分期	T1 期 = 1,T2 期 = 2,T3 期 = 3,T4a 期 = 4,T4b 期 = 5
分化程度	高分化 = 1,中分化 = 2,低分化 = 3
病理学类型	腺癌 = 1,黏液腺癌 = 2,印戒细胞癌 = 3
合并糖尿病	无 = 0,有 = 1
胆囊切除史	无 = 0,有 = 1

2.2　核查等级资料的纳入形式和结果呈现是否完整

在多因素分析过程中,无序多分类资料是以“哑变量”形式纳入;但等级资料的纳入形式,则与无序多分类变量有所区别。等级资料可以“哑变量”或“线性变量”的形式纳入,前者是将等级资料定义为类似“无序多分类”的变量,而后者是将等级资料定义为类似“计量资料”的变量。不同的等级资料纳入形式,结果呈现有所不同。如文献[14]中,T 分期若以哑变量形式纳入,则结果中呈现的数据行应是类别数量减 1,即

有4行结果数据，并需给出对照，若以线性变量纳入（需事先审核线性变量纳入条件满足与否），则结果中呈现的数据行只有1行。这与计量资料一致，系数解释也相应为：每增加/降低1个等级，效应的平均改变量。文献[14]中经编辑后，作者修改结果如表3。由结果可看出：CEA水平和CA-125水平是计量资料，因而只有一行数据结果；T分期是以线性变量纳入模型，因而也只有一行数据结果；分化程度以高分化为对照，以哑变量形式纳入，因而有2行（3－1）数据结果，病理学类型同；合并糖尿病是二分类资料，根据文中解释和赋值表，对照为未合并糖尿病。

表3 多因素非条件logistic回归结果

因 素	β值	SE	Wald χ^2值	OR值	OR 95% CI	P值
CEA水平	1.694	0.626	4.883	5.442	1.169～13.592	0.027
CA-125水平	3.224	0.631	17.028	25.128	3.923～46.534	<0.001
T分期	1.231	0.322	14.618	3.423	1.822～6.443	0.001
中分化	－0.813	0.783	0.000	0.444	0.257～21.323	1.000
低分化	1.219	0.564	4.682	3.384	1.122～10.211	0.001
黏液腺癌	2.640	0.556	22.533	14.017	4.712～41.697	<0.001
印戒细胞癌	1.378	0.999	1.902	3.967	0.560～28.107	0.168
合并糖尿病	0.867	0.683	4.588	2.379	1.132～16.462	0.032

2.3 核查多因素分析结果是否呈现完整

2.3.1 *多重线性回归分析* 笔者认为，多重线性回归分析的呈现结果中，除了常规的回归系数（β值）、标准化偏回归系数、t值和P值外，还应呈现反映模型拟合效果的指标，如R^2值、校正R^2值、剩余标准差等，以及反映变量之间共线性的指标（方差膨胀因子）等。

2.3.2 logistic *回归* logistic回归模型的呈现结果同上述多重线性回归模型，只是需额外呈现OR值、OR值的置信区间及回归系数的Wald χ^2值（用于初步审核模型拟合情况）。对所建立的logistic回归模型，需进行拟合优度检验，常用的检验统计量有剩余差（D）、Pearson χ^2、Hosmer-Lemeshow拟合优度指标，因计算复杂，需通过软件计算，故需由作者提供。

2.3.3 Cox *比例风险回归模型* Cox比例风险回归模型的呈现结果同logistic回归模型，但需注意，统计量是RR值而非OR或HR。Cox比例风险回归模型的诊断也涉及诸多方面，但最重要的是要考虑比例风险假设是否满足，以及自变量间是否存在多重共线性。

2.4 核查回归系数的解释是否符合常规，以及回归系数值和OR/RR值是否对应

在进行结果审查时，需特别注意系数值是否符合专业知识。假设某死亡影响因素分析时，TNM分期的系数值为负值，对照为TNM Ⅰ＋Ⅱ期，则解释为患者TNM Ⅲ＋Ⅳ期的死亡率低于TNM Ⅰ＋Ⅱ期，这和专业认知相悖，需仔细审核模型自变量纳入方法、进行共线性诊断等，排除共线性对TNM分期效果呈现的干扰。

此外，在logistic回归分析和Cox比例风险回归模型条件下，若回归系数的值为负值，则OR/RR值＜1，若回归系数的值为正值，则OR/RR值＞1。编辑也需注意回归系数和OR/RR值的逻辑对应关系。如文献[15]，编修前作者提供的多因素分析结果见表4。表4存在的数据问题包括：1）年龄无OR值可信区间和P值数据；2）脾静脉直径、门静脉直径、术前门静脉血流速度、术后第7天D-二聚体水平及血小板计数变量的95% CI中包含了1，P值却小于0.05；3）术后第7天D-二聚体水平及术后第7天血小板计数变量的回归系数为负值，OR值却大于1；4）门静脉直径的OR值和回归系数不对应。故提请作者审核、修改。

表4 多因素logistic回归分析结果

变 量	β值	Wald值	OR值	95%CI	P值
年龄＞50岁	0.249	5.128	0.02		
脾静脉直径＞12mm	0.523	4.276	1.289	0.861～2.245	<0.01
门静脉直径＞13mm	1.875	6.761	1.551	0.915～1.847	0.01
术前门静脉血流速度＜18 cm/s	0.475	4.182	1.473	0.726～2.215	<0.01
术后第7天D-二聚体水平升高	－6.234	5.231	1.346	0.739～1.896	0.03
术后第7天血小板计数升高	－5.725	4.761	1.691	0.861～2.685	0.02

2.5　核查效应统计量和 95% CI 之间是否对应、核查 95% CI 和 *P* 值是否对应

OR 值、RR 值和各自对应的 95% CI 之间的对应关系为,95%CI 范围内必定包含了 OR/RR 值(图 1)。

OR/RR 值时,涉及 95% CI 和 *P* 值的逻辑对应关系,即若 95% CI 区间包括了 1,则 *P* 值≥0.05,否则 *P* 值 <0.05。编辑人员在编辑稿件时,务必注意两者之间的逻辑关系应一致。如表 4,脾静脉直径、门静脉直径、术前门静脉血流速度、术后第 7 天 D-二聚体水平及血小板计数变量的 95% CI 中包含了 1,*P* 值却小于 0.05,则数据存在错误,需提请作者核查。

2.6　核查标准误或 Wald χ^2 值是否过大

前面提及,Wald χ^2 值过大时,提示变量之间可能存在共线性或模型拟合有问题,需提请作者核查。

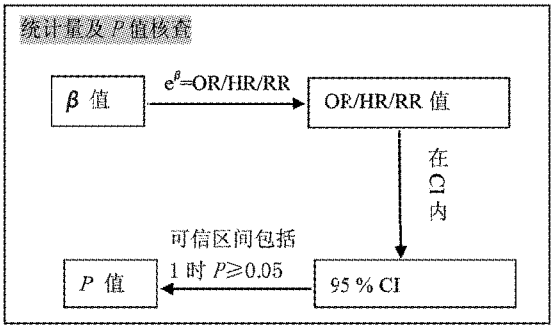

图 1　多因素分析结果的核查要点

3　结束语

医学期刊编辑在工作中经常会遇到采用了多因素分析方法的文章,因而要求编辑能够对多因素分析方法过程及结果进行核查,发现常见错误,给予作者建议,提请作者修改或补充。笔者认为,可操作的解决办法是要求作者提供分析结果原图,这样可以避免一部分数据逻辑错误,如 β 值和 OR/RR 值不匹配,95% CI 和 *P* 值不匹配等问题。

4　参考文献

[1] 李晓松, 陈峰, 郝元涛, 等. 卫生统计学[M]. 8 版. 北京: 人民卫生出版社, 2017: 250

[2] 罗云梅, 蒲素清, 李缨来. 中文医学期刊编辑对生存分析的核查要点[J]. 编辑学报, 2018, 30(1): 32

[3] 孙振球, 徐勇勇. 医学统计学[M]. 2 版. 北京: 人民卫生出版社, 2008: 367

[4] 伍庆林, 沈浩元, 胡超华. 甲状腺乳头状癌颈淋巴结清扫策略的再探讨[J]. 中国普外基础与临床杂志, 2019, 26(12): 1419

[5] 李立强, 谷涌泉, 佟铸, 等. 股动脉内膜剥脱治疗股总动脉硬化闭塞症后再狭窄影响因素的 Cox 比例风险回归分析[J]. 中国普外基础与临床杂志, 2017, 24(9): 1072

[6] 刘雨薇, 徐裕杰, 李卡, 等. 结直肠癌患者术前肠道微生态失衡现状及影响因素分析[J]. 中国普外基础与临床杂志, 2019, 26(10): 1190

[7] 杨梅, 肖静, 蔡辉勇. 多元分析中的多重共线性及其处理方法[J]. 中国卫生统计, 2012, 29(4): 620

[8] 刘国旗. 多重共线性的产生原因及其诊断处理[J]. 合肥工业大学学报(自然科学版), 2001, 24(4): 608

[9] 马雄威. 线性回归方程中多重共线性诊断方法及其实证分析[J]. 华中农业大学学报(社会科学版), 2008(2): 79

[10] 陶然. Logistic 模型多重共线性问题的诊断及改进[J]. 统计与决策, 2008(15): 23

[11] 陈雄飞,董晓梅,汪宁, 等. 多因子共线性的主成分 logistic 回归分析[J]. 中国卫生统计, 2003, 20(4): 213

[12] 张丕德. 预后因子存在共线性时 Cox 模型的拟合方法[J]. 中国卫生统计, 2000, 17(6): 358

[13] 张丕德. Cox 模型多因子共线性处理方法的进一步研究[J]. 中国卫生统计, 2000, 17(4): 207

[14] 朱信强, 管文贤. 结直肠癌同时性腹膜转移影响因素的多因素分析[J]. 中国普外基础与临床杂志, 2016, 23(6): 698

[15] 郑波,廖东旭, 杨训, 等. 门静脉高压性脾功能亢进患者腹腔镜脾切除术后并发门静脉系统血栓形成的危险因素研究[J]. 中国普外基础与临床杂志, 2018, 25(4): 440

(2019-09-23 收稿;2019-12-17 修回)