·论著·方法学研究·

多变量选择方法在临床预测模型中的验证：基于 MIMIC 数据库

郑帅[1,2]，黄韬[1]，杨瑞[1]，李莉[1]，乔萌萌[2]，陈冲[2]，吕军[1]

1. 暨南大学第一附属医院临床研究部（广州 510630）
2. 陕西中医药大学公共卫生学院（陕西咸阳 712046）

【摘要】 **目的** 验证不同变量选择方法对临床预测模型性能的影响。**方法** 从 MIMIC 数据库中提取了 3 组样本数据集（急性心肌梗塞组、脓毒症组和脑出血组），用 COX 回归的直接进入、逐步向前、逐步向后、LASSO、岭回归、基于随机森林的变量重要性六种方法，选出的不同方法的最优变量集构建模型，通过 C 指数、受试者工作特征曲线下面积（AUC 值）和校准曲线，比较组内和组间的结果差异。**结果** 6 种变量选择方法筛选的变量及数目各不相同，但不管是组内还是组间，并没有显示出哪种方法有明显提高模型性能的优势。**结论** 在使用变量选择方法建立临床预测模型前应首先明确研究目的并判断数据的类型，结合医学知识选择合适的方法。

【关键词】 变量选择；临床预测模型；MIMIC 数据库；模型建立

Validation of multivariate selection method in clinical prediction models: based on MIMIC database

ZHENG Shuai[1,2], HUANG Tao[1], YANG Rui[1], LI Li[1], QIAO Mengmeng[2], CHEN Chong[2], LYU Jun[1]

1. Department of Clinical Research, the First Affiliated Hospital of Jinan University, Guangzhou 510630, P.R.China
2. School of Public Health, Shanxi University of Chinese Medicine, Xianyang 712046, P.R.China
Corresponding author: LYU Jun, Email: lyujun2020@jnu.edu.cn

【**Abstract**】 **Objective** To verify the influence of different variable selection methods on the performance of clinical prediction models. **Methods** Three sample sets were extracted from the MIMIC database (acute myocardial infarction group, sepsis group, and cerebral hemorrhage group) using the direct entry of COX regression, step by step forward, step by step backward, LASSO, and ridge regression, based on random forest. These existing six methods of variable importance algorithm, and the optimal variable set of different selected methods were used to construct the model. Through the C index, the area under the ROC curve (AUC value) and the calibration curve, and the results within and between groups were compared. **Results** The variables and numbers selected by the six variable selection methods were different, however, whether it was within or between groups did not reflect which method had the advantage of significantly improving the performance of the model. **Conclusions** Prior to using the variable selection method to establish a clinical prediction model, we should first clarify the research purpose and determine the type of data. Combining medical knowledge to select a method that can meet the data type and simultaneously achieve the research purpose.

【**Key words**】 Variable selection; Clinical prediction model; MIMIC database; Model establishment

在医学研究领域中，观察性研究通常在预后或病因学方面进行，多变量建模是推断流行病学因果关系和调查流行病学预后因素的基本工具[1]。而多元回归模型已广泛应用于健康科学相关的探索性和验证性研究中。

通常，研究数据收集的目的是希望解释某些变量之间存在的相互关系或确定影响特定不良事件的因素。多元回归模型是这类研究中常用的工具，许多观察性研究使用多元回归方法来确定一个结局的重要预测模型[2-4]，最终目标将是获得一个简化的模型。该模型从生物学角度讲是有意义的，而且在应用于独立数据时可提供有效预测[5]。在模型建立之初，应做的是变量选择，即将清洗后的数据，

DOI：10.7507/1672-2531.202107175
基金项目：国家社会科学基金项目（编号：16BGL183）
通信作者：吕军，Email：lyujun2020@jnu.edu.cn

根据相应的科学依据和可靠的统计手段，尽量选择有意义的变量纳入模型进行分析。变量选择有两个目的，首先，它有助于确定与结果相关的变量集合，从而使模型完整，准确。其次，它通过消除不相关的变量来帮助构建紧凑模型，提升模型精度并减少模型复杂性。最终，变量选择应在简单性和完整性之间取得平衡[6]。

常用的变量筛选方法有直接进入法、逐步回归法（向前和向后）、LASSO 法、岭回归和基于随机森林的变量重要性算法等，而如何选择合适的变量筛选方法比较困难。当研究报告了生存时间数据时，COX 模型是最常用的生存分析方法。本文将基于 COX 模型对 MIMIC 数据库中多组数据采用不同变量筛选方法构建临床预测模型，并通过模型的相关性能指数对比进行验证。

1　数据与方法

1.1　数据

本次研究的所有数据均来源于 MIMIC 数据库，MIMIC 是由麻省理工大学计算生理学实验室开发的、可公开获取的数据集，包括约 60 000 例重症监护病房就诊相关的身份不明患者的健康信息数据。数据包括人口统计资料、生命体征、实验室检查和用药方案等。该数据库具有样本量大、数据全面、长期患者追踪、可免费使用等优点，为重症监护研究提供了丰富资源[7,8]。我们从数据库中提取了 3 种疾病的数据作为本次研究的 3 个样本集，样本集 1 是急性心梗患者，包含 4 612 例样本，72 个变量；样本集 2 是脓毒症患者，包含 1 289 例样本，39 个变量；样本集 3 是脑出血患者，包含 813 例样本，76 个变量。3 种疾病研究的结局指标均是死亡，且包含时间协变量，变量中均包含了患者年龄、性别等人口学特征资料，同时也纳入了实验室检查数据。

1.2　方法

对 3 组样本集做基于 COX 比例风险回归的临床预测模型研究。根据不同疾病的特质，分别建立了长-短期生存分析模型。首先我们对 3 组数据进行了描述性分析，其次分别对 3 组样本的分类变量和连续变量进行统计推断，采用卡方检验和 T 检验，得到相关 P 值，可确认根据结局分组在不同变量之间是否存在统计学差异。然后通过 R 语言将数据按 3∶7 的比例划分为训练集和测试集。在筛选变量进入模型时分别采用了直接进入法（P<0.05）、逐步向前法（P<0.05）、逐步向后法（P<0.05）、Lasso 法、岭回归和变量重要性 6 种方法。根据不同的筛选结果各自建立模型，并计算各模型的 C 指数、AUC 值（受试者工作特征曲线下面积）和校准曲线判断模型的性能，从而选择模型性能较好的变量选择方法。数据处理工具包括 R 语言、SPSS、EXCEL。模型建立流程图如图 1 所示。

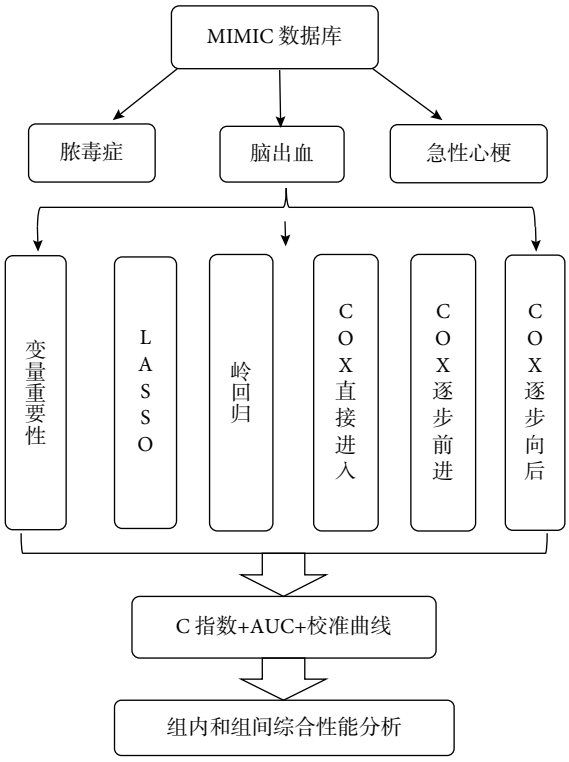

图 1　模型建立的流程图

2　结果

根据结局（生/死）对分类变量和连续变量分别做了卡方检验和 t 检验，并计算每组变量的 P 值，结果如超链接所示（各组患者的变量特征表），3 组数据均以患者发生死亡为结局，其中急性心梗组有 4 612 例患者，纳入了 26 个分类变量和 46 个连续变量；脓毒症组有 1 289 例患者，纳入了 4 个分类变量和 35 个连续变量；而脑出血组有 813 例患者，纳入了 28 个分类变量和 48 个连续变量。各组通过 6 种不同变量筛选方法的统计结果如表 1 所示，在急性心梗组又 12 个变量被 6 种方法共同识别对预后有影响，脓毒症组有 9 个变量，而脑出血组有 6 个变量。但在急性心梗组利用变量重要性的方法可识别 47 个变量，是利用直接进入法识别个数（18 个）的两倍多。而其他两组在利用不同方法识别时变量数量的表现相对稳定。

各组的 C 指数和 AUC 的统计结果如表 2、3 所示，在测试集中急性心梗组 C 指数表现最好的是岭

回归（0.833），AUC 表现最好的是变量重要性（0.819）；脓毒症组 C 指数最高的是逐步向后 LR（0.731），AUC 值最高的是 Lasso（0.754）；而脑出血组 C 指数和 AUC 表现最好的分别是逐步向后 LR（0.770）和直接进入法（0.842），3 组数据中各模型都表现出良好的校准度，预测值与真实值接近，但并没有凸显出某一方法具有独特的优势（校准图可联系作者获取）。

3 讨论

生物医学研究中的因果关系问题需要采用观察性研究和实验性研究等多种试验设计方法去验证。而流行病学研究中多变量建模的结果为评估假定的危险因素在人类疾病中的潜在因果作用提供了宝贵的信息[1]。但在建模之初，变量的有效识别是影响整个模型的真实性和有效性的直接因素。

由于数字化的快速发展，大数据在医疗保健领域已成为重要的数据来源。精准健康包括根据可用的临床和生物学数据应用适当的统计模型，以更准确地预测患者的预后。而大数据集包含数千个变量，这使得传统方法很难有效地处理和管理数据。因此，变量选择已成为大数据分析领域许多研究的重点[9]

从数据集中所有可用变量中识别出潜在候选变量后，进一步选择变量将其包含在最终模型中。选择模型变量有不同的方法，但是关于哪种方法最好目前尚无共识[10]。本文将常见的 6 种变量筛选方法应用于 3 种疾病的生存分析研究，并在表 2 和表 3 中总结了最终模型的性能指标。本次研究的结果显示不同方法识别的变量和数量均有差异，但在 3 个数据集中均未发现哪种方法的使用可使模型的性能总体得到提升，结果同时表明模型纳入的变量数并不是影响模型结果的因素，打破了“包含变量数越多，模型越好”的悖论。在 C 指数和 AUC 值的统计结果中，心梗组和脑出血组的总体表现优于脓毒症组，可看出似乎数据本身才是影响模型性能的根本原因，校准图的表现同样证明了此观点。尽管变量选择方法易于使用且易于构建多变量模型，但从业人员（如数据收集者）常常忽视诸如选择不确定性或报告数量偏差等问题[11]。如何合理地选择统计方法应严格取决于所要解决的研究问题本身，这对模型构建、数据分析和数据解释具有重要影响[1]。

表 1 各组不同方法筛选的变量数

变量选择方法	急性心梗	脓毒症	脑出血
变量重要性	47	17	22
LASSO	18	18	32
岭回归	23	16	19
COX 直接进入	19	15	20
COX 逐步向前	21	14	15
COX 逐步向后	30	17	24
被 6 种方法共同识别	12	9	6

表 2 C 指数的统计结果

变量选择方法	训练集			测试集		
	AUC1	AUC2	AUC3	AUC1	AUC2	AUC3
变量重要性	0.845	0.689	0.753	0.819	0.752	0.788
LASSO	0.840	0.711	0.752	0.806	0.754	0.826
岭回归	0.832	0.711	0.759	0.808	0.745	0.817
COX 直接进入	0.840	0.707	0.747	0.807	0.752	0.842
COX 逐步向前	0.844	0.711	0.761	0.810	0.741	0.810
COX 逐步向后	0.847	0.710	0.763	0.818	0.736	0.827

AUC：受试者工作特征曲线下面积；1：急性心梗；2：脓毒症；3：脑出血。

表 3 AUC 的统计结果

变量选择方法	训练集			测试集		
	C1	C2	C3	C1	C2	C3
变量重要性	0.846	0.728	0.748	0.827	0.716	0.735
LASSO	0.840	0.746	0.754	0.826	0.725	0.755
岭回归	0.844	0.744	0.750	0.833	0.729	0.751
COX 直接进入	0.844	0.740	0.755	0.829	0.713	0.760
COX 逐步向前	0.844	0.741	0.761	0.828	0.729	0.766
COX 逐步向后	0.846	0.742	0.762	0.832	0.731	0.770

AUC：受试者工作特征曲线下面积；C：C 指数；1：急性心梗；2：脓毒症；3：脑出血。

COX 比例风险回归是一种多因素的生存分析方法，它可同时分析众多因素对生存期的影响，且不要求估计资料的生存函数分布类型，因此它的使用范围极其广泛，在处理生存分析数据时非常受欢迎，临床预测模型也多基于此建立。但是 COX 模型要求变量间相互独立（至少不能存在很强的关联），且所研究的样本量要大于变量总数，如果忽略这些条件将会降低模型的稳定性和可解释性[12]。一般在使用 COX 回归时可分别使用 3 种不同的变量进入方法（直接进入法、逐步向前进入法和逐步向后进入法），这 3 种方法是基于 SPSS 完成。岭回归的基本原理也是基于修正后的最小二乘法，它的算法在限定了某些系数后使残差平方和最小化，它的优点是可有效处理多重共线性使模型更加稳定，提高预测性，它的缺点是在处理变量多而样本少的情况下，得到较多的自变量，影响模型的可解释性。LASSO 回归相比岭回归简化了模型，减少了不必要的自变量，提高了模型的预测性能。而 LASSO 的缺点是当自变量远多于样本量时，可能会丢失一些非常重要的有意义的变量，导致回归模型的可信度会降低[13]。LASSO 变量选择方法与常规回归相比在样本量小的研究中更能体现它的优势[14]。而变量重要性是一种新的基于随机森林的特征选择方法，它提供了从信息系统中无偏且稳定地选择重要和非重要属性的方法，本次研究的变量重要性结果是基于 R 语言的 Boruta 包实现（http://CRAN.R-project.org/package=Boruta）。该算法通过比较真实特征与随机特征的相关性来确定相关性，它迭代地删除统计测试证明与随机特征相关性较低的特征，采用了一种新颖的特征选择算法来查找所有相关变量[15]。在模型建立的时候每个步骤都有其参考的指标，如变量纳入模型时一般要求 $P<0.05$，但 P 值同样不能量化模型在预测时犯错的概率，这只是可接受的阈值[16]。在设定 P 值时，应严格按照研究设计方案，根据样本量分析、研究目的的性质以及临床经验设定合理的检验标准（alpha），平衡研究结果中的一类错误和二类错误。从整体模型的性能考虑，常规回归一般会先择 AIC、BIC 值最小的模型，但这也只是在现有方法中选取最好的模型的手段，目前深度学习、机器学习相继提出新的变量选择的算法[17,18]，但同样没有指出哪种变量选择方法是最佳的。

本文的研究结果从区分度和校准度角度验证了上述 6 种变量选择方法在应用于临床预测模型研究的数据时，并没有哪一种方法明显使模型的性能提高。总之，在使用变量选择方法之前，应该批判性地考虑在特定研究中是否完全需要这样的方法，如果是，则仅通过“让数据说话”就有足够的理由来证明在模型中消除或包含变量是合理的[11]。因此，建模应基于有背景知识和可验证的假设开始，这些知识应来自于同一研究领域的前期研究、专家经验或常识。遵循这一黄金法则，通常可在不使用手头数据集来揭示变量与结果之间的关系的情况下就可建立一个初始的变量集合，即“全模型”[19]，但是更多情况下，研究的目的是探索一些未知的变量对结局的影响，因此变量选择是研究中必不可少的。而样本大小和候选预测变量个数是最有影响力的模拟条件[20]。试图以选取某一变量选择方法或盲目增加变量来提高模型性能的手段是不可靠的。而收集到高质量的数据是建立一个拥有优良性能模型的前提，其次通过合理的研究设计，在合成数据集时就应依据文献指南等知识排除混杂因素，不能以“丰富数据”或“探索未知”的思维而纳入各种不确定因素导致结论与真实情况的偏倚增大，而是应结合研究目的，以最终数据类型选取合适的变量选择方法。

本文的局限性：① 此研究仅涉及到半参数 COX 回归的生存模型建立；② 数据仅来源于重症相关疾病，在后续的研究中还需利用更丰富的数据以及更多的变量选择方法验证此结论，以提高结论的普适性。

总之，在使用变量选择方法建立临床预测模型前应首先明确研究目的并判断数据类型，结合医学知识选择可同时满足数据类型和达到研究目的的方法。

参考文献

1 Tripepi G, Jager KJ, Dekker FW, *et al*. Testing for causality and prognosis: etiological and prognostic models. Kidney Int, 2008, 74(12): 1512-1515.

2 von Düring ME, Jenssen T, Bollerslev J, *et al*. Visceral fat is better related to impaired glucose metabolism than body mass index after kidney transplantation. Transpl Int, 2015, 28(10): 1162-1171.

3 Bhat M, Hathcock M, Kremers WK, *et al*. Portal vein encasement predicts neoadjuvant therapy response in liver transplantation for perihilar cholangiocarcinoma protocol. Transpl Int, 2015, 28(12): 1383-1391.

4 Pianta TJ, Peake PW, Pickering JW, *et al*. Evaluation of biomarkers of cell cycle arrest and inflammation in prediction of dialysis or recovery after kidney transplantation. Transpl Int, 2015, 28(12): 1392-1404.

5 Núñez E, Steyerberg EW, Núñez J. Regression modeling strategies. Rev Esp Cardiol, 2011, 64(6): 501-507.

6 Chowdhury MZI, Turin TC. Variable selection strategies and its

importance in clinical prediction modelling. Fam Med Community Health, 2020, 8(1): e000262.
7 Johnson AE, Pollard TJ, Shen L, *et al*. MIMIC-Ⅲ, a freely accessible critical care database. Sci Data, 2016, 3: 160035.
8 Yang J, Li Y, Liu Q, *et al*. Brief introduction of medical database and data mining technology in big data era. J Evid Based Med, 2020, 13(1): 57-69.
9 Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res, 2003, 3(3): 1157-1182.
10 Royston P, Moons KG, Altman DG, *et al*. Prognosis and prognostic research: Developing a prognostic model. BMJ, 2009, 338: b604.
11 Heinze G, Dunkler D. Five myths about variable selection. Transpl Int, 2017, 30(1): 6-10.
12 Zhang Z, Reinikainen J, Adeleke KA, *et al*. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med, 2018, 6(7): 121.
13 Alves LF, Fernandes BF, Burnier JV, *et al*. Incidence of epithelial lesions of the conjunctiva in a review of 12 102 specimens in Canada (Quebec). Arq Bras Oftalmol, 2011, 74(1): 21-23.
14 张玉. 自变量个数远大于样本数情形下($p>n$)罚函数回归法的改进. 江苏教育学院学报(自然科学版), 2012, 28(3): 28-32.
15 Kursa M, Rudnicki W. Feature selection with the boruta package. J Stat Soft, 2010, 36: 1-13.
16 Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol, 2008, 45(3): 135-140.
17 Han J, Zheng H, Xing Y, *et al*. V2V: A deep learning approach to variable-to-variable selection and translation for multivariate time-varying data. IEEE Trans Vis Comput Graph, 2021, 27(2): 1290-1300.
18 Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform, 2019, 20(2): 492-503.
19 Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom J, 2018, 60(3): 431-449.
20 Wiegand RE. Performance of using multiple stepwise algorithms for variable selection. Stat Med, 2010, 29(15): 1647-1659.

收稿日期：2021-07-26 修回日期：2021-11-03
本文编辑：熊鹰