DOI:10.12138/j.issn.1671－9638.20195051

·论著·

Lasso-logistic 模型在医院下呼吸道感染预测中的应用

康文博[1],赵静雅[2],吕雪峰[3],陈　勇[2],韩雪琳[2],田曙光[2],陈芳艳[2],苏雪婷[2],王洪源[1],韩　黎[2]
(1. 北京大学公共卫生学院,北京　100191; 2. 中国人民解放军疾病预防控制中心医院感染监控中心,北京　100071; 3. 中央军委后勤保障部信息中心,北京　100842)

[摘　要] **目的**　建立住院患者医院下呼吸道感染预测模型,构建新的、简单的风险评分方法。**方法**　以 2014 年多家医院感染调查数据为训练集,建立住院患者医院下呼吸道感染的 Lasso-logistic 回归预测模型,选择贝叶斯信息准则(BIC)最小模型为最终模型,将回归系数放大相同倍数建立评分方法,以 2015、2016 年调查数据为验证集,并与文献建立的风险评分方法进行比较。**结果**　Lasso 过程共进行 360 步,第 24 步时 BIC 最小(6 690.4),正则化参数 $\lambda=130.8$。风险评分方法包含 17 个条目,数量是文献风险评分方法的 1/4,DeLong's 检验显示,两评分方法验证集受试者工作特征曲线下面积(AUC)差异无统计学意义($Z=0.371$,$P=0.710$),决策曲线几乎重合,净重新分类指数为 −0.0149,差异无统计学意义($Z=-1.301$,$P=0.193$),整体鉴别指数为 0.006,改善差异有统计学意义($P=0.014$)。**结论**　利用 Lasso-logistic 回归模型建立了住院患者医院下呼吸道感染风险简单评分方法,该方法的条目相对简洁,预测效果准确。

[关　键　词] Lasso; 医院感染; 下呼吸道感染; 风险评分; 预测
[中图分类号] R181.3+2

Application of Lasso-logistic model in prediction of healthcare-associated lower respiratory tract infection

KANG Wen-bo[1], *ZHAO Jing-ya*[2], *LV Xue-feng*[3], *CHEN Yong*[2], *HAN Xue-lin*[2], *TIAN Shu-guang*[2], *CHEN Fang-yan*[2], *SU Xue-ting*[2], *WANG Hong-yuan*[1], *HAN Li*[2] (1. *School of Public Health, Peking University, Beijing* 100191, *China*; 2. *Center for Healthcare-associated Infection Surveillance and Control, Chinese PLA Center for Disease Control and Prevention, Beijing* 100071, *China*; 3. *Information Center of the Logistics Support Department of the Central Military Commission, Beijing* 100842, *China*)

[**Abstract**] **Objective**　To develop a predictive model for healthcare-associated lower respiratory tract infection (HA-LRTI) in hospitalized patients, and establish a simple risk scoring method. **Methods**　Survey data of healthcare-associated infection(HAI)in a few hospitals in 2014 was as training dataset, a Lasso-logistic regression model for predicting HA-LRTI in hospitalized patients was established, minimum model of Bayesian information criterion (BIC) was chosen as the final model, scoring method was established by magnifying regression coefficient by the same scale, survey data of 2015 and 2016 were used as the validation dataset, and was compared with risk scoring method established in the literatures. **Results**　Among the 360 steps of Lasso, smallest BIC (6 690.4) occurred at step 24 with regularization parameter $\lambda=130.8$. The risk scoring method consisted 17 items, which was 1/4 of the amount of literature risk scoring method, DeLong's test showed that there was no significant difference in area under the curve of receiver operating characteristic between two scoring methods ($Z=0.371$, $P=0.710$), decision curve analysis almost overlaid, the net reclassification index was −0.0149, with no significant difference ($Z=$

[收稿日期]　2019－01－15
[基金项目]　国家科技重大专项(2018ZX10733402; 2018ZX10713003)
[作者简介]　康文博(1994－),女(汉族),河南省周口市人,硕士研究生,主要从事医疗数据分析。
[通信作者]　王洪源　　E-mail:why_w2003@163.com

－1.301，$P=0.193$），the integrated discrimination index was 0.006，and difference was significant（$P=0.014$）.

Conclusion Lasso-logistic regression model established a simple scoring method of HA-LRTI risk for inpatients，the items of the method is relatively concise and the predictive effect is accurate.

［**Key words**］ Lasso；healthcare-associated infection；lower respiratory tract；risk score；prediction

据世界卫生组织（WHO）2011 年估计，世界范围内每年有上亿人受到医院感染的影响，医院感染已经成为一个严重的全球公共卫生问题，中低收入水平国家的医院感染负担远高于高收入水平国家（医院感染现患率分别为 15.5% 和 7.6%）[1]。2008—2014 年全国医院感染监测网横断面调查结果显示，我国医院感染现患率逐渐下降，感染类型以下呼吸道感染为主[2-5]。准确的临床预测模型可以帮助筛选医院感染的高危对象，提高医院感染防控措施的针对性和效率。Chen 等[6]利用 2014 年一项多所医院医院感染横断面调查数据，构建了医院下呼吸道感染风险评分方法（以下称为原始评分方法），共包括 70 个条目，训练集回代预测效果较好，但相对复杂的预测方法可能不利于临床日常使用[7]。Tibshirani[8]在 1996 年提出了 Lasso（Least absolute shrinkage and selection operator），通过 L1 惩罚对自变量回归系数进行压缩，可以将对模型影响较小的变量系数压缩为 0，筛选出相对重要的变量。使用 Lasso 方法的 logistic 回归又叫做 Lasso-logistic 回归，已有研究者将其应用于出生缺陷[9]、老年痴呆[10]等医学研究领域，其中李敏捷[9]建立的 Lasso-logistic 回归出生缺陷预测模型效果好于逐步法得到的 logistic 回归模型。本文以 2014 年调查数据为训练集，建立精简的医院下呼吸道感染 Lasso-logistic 回归预测模型，构建新的风险评分方法，并以 2015、2016 年调查数据为验证数据，与原始评分方法进行比较。

1 对象与方法

1.1 *数据来源* 研究数据来源于一项多所医院医院感染联网监测横断面调查，2014—2016 年每年调查一次，其中 2014 年调查患者 52 561 例，2015 年 30 313 例，2016 年 26 320 例。调查内容包括住院患者的一般情况、基础疾病状况、住院期间治疗和医院感染发生情况。

1.2 *医院感染诊断标准* 依据卫生部《医院感染诊断标准（试行）》（卫医发［2001］2 号）[11]进行医院感染诊断。

1.3 *研究对象特征描述* 描述训练集与验证集纳入研究对象的一般特征、变量赋值见表 1。

表 1 研究对象特征描述变量赋值表

Table 1 Variable assignments of characteristic description of research objects

分类变量	赋值情况
性别	0＝女，1＝男
泌尿道插管	0＝否，1＝是
中央或周围动静脉置管	0＝否，1＝是
使用呼吸机	0＝否，1＝是
气管切开	0＝否，1＝是
血液透析	0＝否，1＝是
使用抗菌药物	0＝否，1＝是
手术切口类型	未手术＝0，Ⅰ类切口＝1，Ⅱ类切口＝2，Ⅲ类切口＝3，Ⅳ类切口＝4
患有 ICD10 类目对应疾病	0＝否，1＝是
医院下呼吸道感染	0＝否，1＝是

1.4 *Lasso-logistic 回归预测模型的建立* 以医院下呼吸道感染诊断情况为结局变量，共纳入自变量 247 个，根据贝叶斯信息准则（Bayesian information criterion，BIC）选择合适的正则化参数 λ，回归系数非 0 的变量纳入最终模型。Lasso 的 L1 正则化路径估计使用预测－校正法（predictor-corrector method），各变量回归系数扩大相同的倍数后四舍五入取整，作为新的住院患者医院下呼吸道感染风险评分的风险指数。

1.5 *预测效果评价* 训练集的预测效果评价使用回代法。预测效果的评价采用受试者工作特征（receiver operating characteristic，ROC）曲线，灵敏度和特异度、阳性似然比和阴性似然比，以及净重新分类指数（net reclassification index，NRI）、整体鉴别指数（integrated discrimination index，IDI）和决策曲线（decision curve analysis，DCA）。

1.6　统计学处理　主要应用 R(3.4.0)和 SAS(9.4)软件进行统计分析，其中 Lasso-logistic 回归模型的建立使用 R 软件的 glmpath 包。不同 ROC 曲线间的比较常用的指标为 ROC 曲线下面积(area under curve，AUC)[12]，对于配对 ROC 曲线，很小的 AUC 差别也可能是有统计学意义的[13]，采用 DeLong's 检验比较不同评分方法在验证集的预测 AUC，检验水准取 $\alpha=0.05$。

2　结果

2.1　一般特征　训练集共纳入研究对象 49 328 例，其中 839 例发生医院下呼吸道感染，发病率为 1.7%；验证集纳入研究对象 50 997 例，其中 783 例发生医院下呼吸道感染，发病率为 1.5%。验证集人群男性比例、住院期间接受各种侵入性操作的比例均高于训练集，使用抗菌药物的比例低于训练集，其他特征相近。见表 2。

表 2　训练集与验证集研究对象的一般特征[例(%)]

Table 2　General characteristics of research objects of training dataset and validation dataset (No. of cases [%])

变量	训练集 ($n=49\ 328$)	验证集 ($n=50\ 997$)
年龄[岁，中位数(P_{25}，P_{75})]	52.0 (35.0,66.0)	53.3 (36.0,67.3)
住院周数[中位数(P_{25}，P_{75})]	1.14 (0.57,2.14)	1.14 (0.57,2.00)
性别		
男性	21 173(42.9)	30 933(60.7)
女性	28 155(57.1)	20 064(39.3)
泌尿道插管	6 801(13.8)	8 249(16.2)
动静脉插管	4 239(8.6)	6 319(12.4)
使用呼吸机	2 272(4.6)	2 929(5.7)
气管切开	814(1.7)	922(1.8)
血液透析	615(1.2)	910(1.8)
使用抗菌药物	7 191(14.6)	6 903(13.5)
手术	10 518(21.3)	13 457(26.4)
医院下呼吸道感染	839(1.7)	783(1.5)

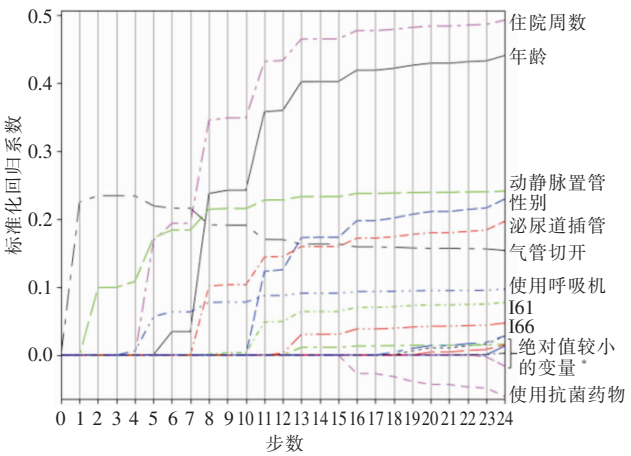
2.2　Lasso-logistic 回归与简单评分　Lasso 过程共进行了 360 步，初始正则化参数 λ_{max} 为 1 335.6。第 24 步时 BIC 达到最小值 6 690.4，$\lambda=130.8$，模型中非 0 回归系数有 17 个，参数估计结果见表 3。使用抗菌药物、手术切口清洁度高的患者医院下呼吸道感染风险降低，其他变量均为患者医院下呼吸道感染的危险因素。最先"进入"模型(回归系数在某步后变为非 0)的变量依次是气管切开和动静脉置管。年龄和住院时间对患者医院下呼吸道感染风险影响明显高于其他变量，见图 1。以年龄回归系数的绝对值为 1 个单位，各回归系数除以该值后四舍五入成整数作为风险指数，构建风险评分方法(见表 4)，如性别的回归系数为 0.463，0.463/0.142≈3，则风险指数为 3。简单评分的训练集 AUC 为 0.883 [95%CI (0.872,0.895)]，推荐以 14 分为预测分割点，灵敏度和特异度分别为 0.84、0.76，阳性似然比和阴性似然比分别为 3.54、0.21。

表 3　Lasso-logistic 回归最终模型参数估计

Table 3　Estimated parameters of final Lasso-logistic regression model

变量	回归系数	标准化系数
年龄*	0.142	0.441
性别	0.463	0.229
住院周数*	0.400	0.493
动静脉置管	0.860	0.241
泌尿道插管	0.570	0.197
气管切开	1.207	0.154
手术切口类型	−0.020	−0.016
使用呼吸机	0.462	0.097
使用抗菌药物	−0.171	−0.061
ICD10 疾病类目		
支气管和肺恶性肿瘤(C34)	0.171	0.029
髓样白血病(C92)	0.211	0.014
颅内出血(I61)	0.684	0.078
大脑动脉闭塞和狭窄无脑梗死(I66)	1.655	0.047
气管和支气管先天畸形(Q32)	3.480	0.016
其他协调缺乏(R27)	1.319	0.017
其他的一般症状和体征(R68)	0.129	0.003
器官和组织移植状态(Z94)	0.391	0.029

*：模型中使用的年龄和住院时间变量非原始变量，均根据由限制性立方样条(restricted cubic spline，RCS)得到的非线性相关关系进行了重新赋值，年龄(岁)赋值规则如下：[0,5]=2，(5,15]=1，(15,35]=0，(35,40)=1，40 岁以上每 5 岁一个组(含下限不含上限)依次加 1；住院时间(周)的赋值规则如下：[0,1)=0，[1,2)=1，[2,3)=2，[3,4)=3，≥4=4，重新赋值后的变量均作为连续变量纳入模型



*：标准化回归系数绝对值较小的变量，从上至下依次为 Z94、C34、R27、Q32、C92、R68、手术切口类型

图 1　最终模型内变量 0～24 步标准化回归系数路径图

Figure 1　Standardized regression coefficient path of step 0－24 for variables included in final model

表 4　医院下呼吸道感染患者风险简单评分表

Table 4　Simple risk scoring system for healthcare-associated lower respiratory tract infection

项目	风险指数*
接受Ⅳ类切口手术的患者	－1
使用抗菌药物	－1
年龄（岁）：[0，5]得 2， (5，15]得 1， (15，35]得 0 (35，40)得 1 ≥40 每 5 岁一个组（含下限不含上限），依次加 1	1
患有 C34、C92、R68 疾病	1
男性	3
使用呼吸机	3
住院时间每满一周加 3，最多加 12	3
患有 Z94 病症	3
泌尿道插管	4
患有 I61 疾病	5
动静脉置管	6
气管切开	9
患有 R27 疾病	9
患有 I66 疾病	12
患有 Q32 疾病	25

*：研究对象风险得分左侧项目对应风险指数值总和值

2.3　预测效果比较　简单评分与原始评分方法评分的验证集 ROC 曲线几乎重合，DeLong's 检验显示 AUC 差异无统计学意义（Z = 0.371，P = 0.710），见图 2。在推荐分割点 14 分处，简单评分的灵敏度和特异度分别为 0.84、0.76，阳性似然比和阴性似然比分别为 3.54、0.21。两评分的决策曲线几乎重合，见图 3。阈概率在[0，0.2]时，两种评分的净收益均明显高于 None 模型；当阈概率大于 0.2 时，与 None 模型无明显差别，无应用价值。依据推荐预测分值为阈值（原始评分方法及本研究提出的简单评分中均推荐 14 分为预测分割点）建立预测结果的重分类表（见表 5），计算简单评分相比于原始评分方法的 NRI 值为－0.0149，说明净重新分类收益无统计学意义（Z = －1.301，P = 0.193），IDI 值 0.006，95% CI 为（0.001，0.010），说明整体鉴别的改善有统计学意义（P = 0.014）。

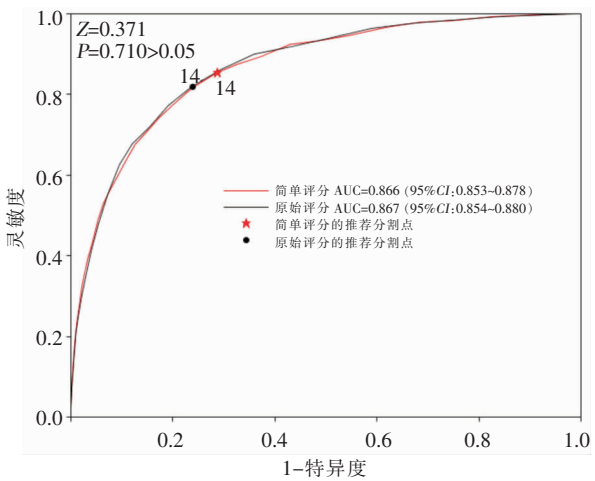


图 2　简单评分与原始评分方法的验证集 ROC 曲线

Figure 2　ROC curves of simple and original scoring methods in validation dataset

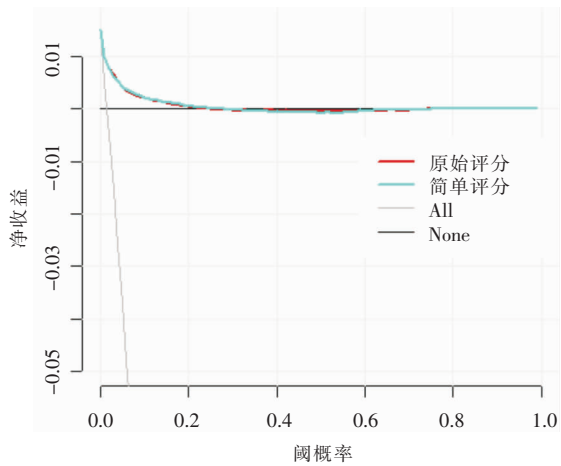


图 3　简单评分与原始评分方法的决策曲线

Figure 3　Decision curves of simple and original scoring methods

表 5　原始评分与简单评分方法的预测结果重分类表

Table 5　Reclassification of predicative result of original and simple scoring methods

原始评分方法	简单评分					
	患者组			非患者组		
	阴性	阳性	合计	阴性	阳性	合计
阴性	89	53	142	34 035	4 213	38 248
阳性	26	615	641	1 732	10 234	11 966
合计	115	668	783	35 767	14 447	50 214

3　讨论

Lasso-logistic 回归预测模型显示，住院患者医院下呼吸道感染的发生受人口学特征、基础疾病特征及住院时间和接受治疗情况的影响，与以往研究[14-15]结果较一致，其中患者的住院日数、年龄对医院下呼吸道感染的影响较大，性别、侵入性操作、使用抗菌药物的影响属于中等水平，手术切口类型及各类基础疾病的影响相对较小，说明在医院下呼吸道感染的防控工作中，应重点关注住院时间较长的高龄、男性患者，规范侵入性操作前、中、后的感染预防措施。

Chen 等[6]构建的住院患者医院下呼吸道感染的风险评分方法包括 70 个条目。本研究建立的简单评分方法仅包含 17 个条目，数量上减少＞3/4，且验证集的预测效果相近，是对原始评分方法的一次成功简化。评分条目的减少主要表现在基础疾病方面，原始评分方法中包括了 61 个 ICD10 类目，简单评分中仅包含 8 个 ICD10 类目，其中风险指数较高的疾病类目包括颅内出血(I61)、大脑动脉闭塞和狭窄无脑梗死(I66)、气管和支气管先天畸形(Q32)、其他协调缺乏(R27)，对患有以上疾病类目对应疾病的住院患者护理工作应该得到加强。原始评分方法中，结肠恶性肿瘤(ICD10 类目为 C18)、前列腺增生(ICD10 类目为 N40)的风险指数为－8，说明患有这些疾病的患者医院下呼吸道感染的风险低于非此类疾病的患者，通常疾病会使机体免疫力下降，简化后评分中各类基础疾病的风险指数均为正值，从免疫学角度上可能更合理，说明 Lasso 过程确实剔除了一些噪声变量。

简单评分的训练集 AUC 为 0.883，验证集 AUC 为 0.866，优于其他医院感染预测研究[16-20]，判别能力良好，与训练集相比验证集 AUC 仅下降了 0.017，与其他验证研究相比下降幅度较小[18-19]，预测效果稳定。

Lasso 的变量压缩程度及预测效果依赖于正则化参数的选择。国内研究者[9,21-23]将 Lasso 应用于健康领域相关研究时，多使用交叉验证选择最终模型。实际上，使用 BIC 选择正则化参数 ，可以得到与真实模型高度一致的变量选择结果，当自变量中的噪声变量较多时，BIC 能够在预测误差相对小的前提下，选择出更为简练的模型[24-25]。本研究首次将 Lasso-logistic 回归模型应用于医院感染研究中，根据 BIC 准则选择正则化参数达到了预期效果，在挑选出更少、更重要自变量的同时保证了预测的准确性，可以为研究者使用 Lasso 方法选择合适的正则化参数选择策略提供经验。

本研究可能存在以下局限性：医院感染的发生除与住院患者自身情况、医疗干预有关外，还可能受到医院的微生物环境等因素的影响，如某科室病房内有患者近期发生过医院感染，则提示环境中可能存在某种易感微生物，此时住院患者感染的风险可能会增加。本研究使用的调查数据不包含医院微生物环境方面的信息，如果纳入相关的变量，预测效果可能会进一步提高。

[参 考 文 献]

[1] Word Health Organization. The burden of health care-associated infection worldwide[EB/OL]. [2018－08－28]. http://www.who.int/gpsc/country_work/burden_hcai/en/.

[2] 任南，文细毛，吴安华. 2008 年全国医院感染横断面调查报告[C]. 中国医院协会全国医院感染管理学术年会，2009：83－87.

[3] 文细毛，任南，吴安华. 2010 年全国医院感染横断面调查感染病例病原分布及其耐药性[J]. 中国感染控制杂志，2012，11(1)：1－6.

[4] 吴安华，文细毛，李春辉，等. 2012 年全国医院感染现患率与横断面抗菌药物使用率调查报告[J]. 中国感染控制杂志，2014，13(1)：8－15.

[5] 任南，文细毛，吴安华. 2014 年全国医院感染横断面调查报告[J]. 中国感染控制杂志，2016，15(2)：83－87.

[6] Chen Y，Shan X，Zhao J，et al. Predicting nosocomial lower respiratory tract infections by a risk index based system[J]. Sci Rep，2017，7(1)：15933.

[7] Moons KG，Royston P，Vergouwe Y，et al. Prognosis and prognostic research：what，why，and how? [J]. BMJ，2009，338：b375.

[8] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. J R Statist Soc，1996，58(1)：267－288.

[9] 李敏捷. Lasso-Logistic 与 Group Lasso-Logistic 模型在出生缺陷研究中的应用[D]. 山西:山西医科大学, 2016.

[10] Lee SH, Yu D, Bachman AH, et al. Application of fused Lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease[J]. J Neurosci Methods, 2014, 221: 78 - 84.

[11] 中华人民共和国卫生部. 医院感染诊断标准(试行)[J]. 中华医学杂志, 2001, 81(5): 460 - 465.

[12] Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach[J]. Biometrics, 1988, 44(3): 837 - 845.

[13] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves[J]. Bmc Bioinformatics, 2011, 12: 77.

[14] 许林勇, 张延霞, 田晓丽, 等. 医院感染危险度的预测预报[J]. 中国医师杂志, 2004, 6(2): 165 - 167.

[15] Kofteridis DP, Papadakis JA, Bouros D, et al. Nosocomial lower respiratory tract infections: prevalence and risk factors in 14 Greek hospitals[J]. Eur J Clin Microbiol Infect Dis, 2004, 23(12): 888 - 891.

[16] 谢多双, 来瑞平, 符湘云,等. ICU 患者医院感染 logistic 回归模型预测[J]. 中华医院感染学杂志, 2011, 21(12): 2424 - 2426.

[17] 陈立兵, 刘运喜, 杜明梅, 等. BP 神经网络在预测血液病患者医院感染中的应用[J]. 中华医院感染学杂志, 2014, 24(6): 1542 - 1544.

[18] Sanagou M, Wolfe R, Leder K, et al. External validation and updating of a prediction model for nosocomial pneumonia after coronary artery bypass graft surgery[J]. Epidemiol Infect, 2014, 142(3): 540 - 544.

[19] Mahieu M, De Dooy JJ, Cossey VR, et al. Internal and external validation of the NOSEP prediction score for nosocomial sepsis in neonates[J]. Crit Care Med, 2002, 30(7): 1459 - 1466.

[20] 樊虹雨. 老年住院患者医院感染风险预警模式的构建[D]. 山西:山西医科大学, 2017.

[21] 成娟, 梁轩, 郑森爽, 等. 基于 Lasso Logistic 回归模型的乳腺癌高风险人群筛查利用相关因素研究[J]. 中华疾病控制杂志, 2018, 22(6): 551 - 559.

[22] 任晓炜. 基于 Group LASSO 的 Logistic 回归在胰岛素心理抵抗因素分析中的应用[D]. 北京:中国人民大学, 2012.

[23] 韩耀风, 覃文峰, 陈炜, 等. adaptive Lasso logistic 回归模型应用于老年人养老意愿影响因素研究的探讨[J]. 中国卫生统计, 2017, 34(1): 18 - 22.

[24] Kirkland LA, Kanfer F, Millard S. Lasso tuning parameter selection[C]. Proceedings of the 57th Annual Conference of SASA, 2015: 49 - 56.

[25] Sun W, Wang J, Fang Y. Consistent selection of tuning parameters via variable selection stability[J]. J Machine Learning Res, 2013, 14(9): 3419 - 3440.

(本文编辑:文细毛)

本文引用格式:康文博,赵静雅,吕雪峰,等. Lasso-logistic 模型在医院下呼吸道感染预测中的应用[J]. 中国感染控制杂志,2019,18(7):619 - 624. DOI:10.12138/j.issn.1671 - 9638.20195051.

Cite this article as: KANG Wen-bo, ZHAO Jing-ya, LV Xue-feng, et al. Application of Lasso-logistic model in prediction of healthcare-associated lower respiratory tract infection[J]. Chin J Infect Control, 2019, 18(7): 619 - 624. DOI:10.12138/j.issn.1671 - 9638.20195051.