论著·临床研究 doi:10.3969/j.issn.1671-8348.2022.03.007
网络首发 https://kns.cnki.net/kcms/detail/50.1097.R.20211231.1919.008.html(2022-01-04)

基于随机森林算法预测肾病综合征患者的心血管风险*

邹新亮[1],郑万香[1],何国祥[1,2],景 涛[1△]
(1.陆军军医大学第一附属医院心血管内科,重庆 400038;2.贵黔国际总医院心血管内科,贵阳 550000)

[摘要] **目的** 评估随机森林模型对肾病综合征(NS)患者5年心血管疾病风险的预测价值。**方法** 选取陆军军医大学第一附属医院就诊的350例NS患者随访5年的诊疗资料,按照约7∶3的比例划分为训练集和测试集。模型纳入28个预测变量,通过训练集进行随机森林模型构建,测试集数据进行模型验证,选取最优节点值和决策树数目,观察变量的预测重要性并评价模型预测性能。**结果** 随机森林模型最佳节点值为6、最佳决策树数目为446。模型中预测因子重要性排序依次为:肾小球滤过率(eGFR)、年龄、高密度脂蛋白胆固醇(HDL-C)、载脂蛋白B(apoB)、清蛋白(ALB)、载脂蛋白A1(apoA1)、纤维蛋白原(Fib)、血尿酸(UA)、低密度脂蛋白胆固醇(LDL-C)。模型预测的准确率为0.919、精确率为0.935、召回率为0.829,AUC及95%*CI*为0.899(0.832~0.966)。**结论** 随机森林分类算法确定的重要预测因子可为预测NS患者5年心血管风险提供有用的信息,模型预测性能良好。

[关键词] 心血管风险;肾病综合征;随机森林;预测模型
[中图法分类号] R541 [文献标识码] A [文章编号] 1671-8348(2022)03-0393-05

Prediction of cardiovascular risk in patients with nephrotic syndrome based on random forest algorithm*

ZOU Xinliang[1],*ZHENG Wanxiang*[1],*HE Guoxiang*[1,2],*JING Tao*[1△]
(1.*Department of Cardiology,the Southwest Hospital of Army Medical University,Chongqing* 400038,*China*;2.*Department of Cardiology,Guiqian International General Hospital,Guiyang,Guizhou* 550000,*China*)

[**Abstract**] **Objective** To evaluate the value of the random forest model in predicting the risk of cardiovascular disease in patients with nephrotic syndrome (NS) over five years. **Methods** The medical data of 350 patients with NS who were followed up for five years in the hospital were collected. The patients were divided into the train set and the test set according to the ratio of nearly 7∶3. A total of 28 predictive variables were incorporated into the model,the train set was used for the random forest model construction,and the test set for model verification. The optimal node values and the number of decision-making tree were selected to observe the predictive importance of variables and evaluate the model's predictive performance. **Results** The optimal node values and the number of decision-making tree of the random forest model were 6 and 446. The order of importance of predictors in this model was glomerular filtration rate (eGFR),age,high-density lipoprotein cholesterol (HDL-C),apolipoprotein B (apoB),albumins (ALB),apolipoprotein A1 (apoA1),Fibrinogen (Fib),uric acid (UA),low-density lipoprotein cholesterol (LDL-C). The accuracy rate of the random forest model was 0.919,the precision rate was 0.935,the recall rate was 0.829,and the AUC and confidence interval was 0.899 (0.832—0.966). **Conclusion** The important predictors determined by the random forest classification algorithm may provide helpful information for predicting the five -year cardiovascular risk of the NS patients. The model has good predictive performance.

[**Key words**] cardiovascular risk;nephrotic syndrome;random forest;predictive model

* **基金项目**:重庆市科卫联合医学科研重点项目(2022ZDXM005);重庆市卫生适宜技术推广项目(2018jstg036);重庆市研究生科研创新项目(CYS19371)。 **作者简介**:邹新亮(1992—),住院医师,本科,主要从事冠心病的发病机制与防治方面的研究。 △ **通信作者**,E-mail:xnkjt@sohu.com。

肾病综合征(NS)以大量蛋白尿、低蛋白血症以及不同程度的水肿为主要特征，常并发高脂血症和(或)静脉血栓等[1]。NS 患者的心血管风险升高，据研究统计，原发性 NS 患者 5 年心血管事件累积发生率约在 6.1%～8.8%[2-4]。尽管 NS 的人口发病率约为 3/10 万人年[2]，但在如此庞大的人口基数下，NS 罹患心血管疾病的患者数量仍然非常多，给患者家庭和社会造成极大的医疗负担。因此，早期对 NS 患者出现心血管疾病风险进行预测和干预极其重要。针对真实世界中 NS 患者可能存在错综复杂的心血管危险因素，本研究采用机器学习算法中的随机森林模型，对 NS 患者 5 年心血管疾病风险进行预测，现将结果报道如下。

1 资料与方法

1.1 一般资料

本研究为单中心回顾性巢式病例对照研究，收集并选取 1999 年 1 月 1 日至 2014 年 11 月 30 日陆军军医大学第一附属医院就诊的 NS 患者随访 5 年的诊疗资料，以评估和预测 NS 患者心血管风险。研究纳入随访期间 18～85 岁的确诊心血管疾病患者 115 例，并将患者队列中根据性别、年龄、指标时间按照约 1∶2 比例匹配，纳入 235 例无心血管疾病对照者，总计 350 例。将全部患者应用统计学软件算法按照约 7∶3 的比例划分为训练集和测试集。本研究经陆军军医大学第一附属医院伦理委员会批准(批准文号：KY2019153)。

纳入标准：原发性肾病综合征(微小病变肾病、系膜增生性肾小球肾炎、局灶节段性肾小球硬化、膜增生性肾小球肾炎、膜性肾病)或继发性肾病综合征(过敏性紫癜性肾炎和狼疮肾炎)[1]；所有肾病综合征患者均根据活检确诊；没有性别或医疗限制；本研究预测结局包括的心血管疾病为：稳定型冠状动脉疾病、非致命性心肌梗死、不稳定性心绞痛和心血管死亡；其中心血管疾病诊断均有影像学证据支持。

排除标准：诊断为高血压肾病或糖尿病肾病；诊断为急性肾损伤；应用透析治疗的慢性肾病；检测到肾小球滤过率(eGFR)＜45 mL·min^{-1}·1.73 m^{-2} 1 次；第 1 次就诊时已确诊为心血管疾病；非心血管死亡；丢失随访或丢失医疗记录。

1.2 方法

1.2.1 数据收集

从所有患者的医疗记录中收集数据，包括以下变量信息：一般情况，性别、年龄、体重指数(BMI)、民族、吸烟状况、饮酒状况；既往病——外周动脉粥样硬化、血糖升高、高血压病、静脉血栓疾病；药物使用情况，抗血小板药、抗凝药、人血清蛋白(ALB)、血管紧张素Ⅱ受体阻滞剂(ARB)、血管紧张素转化酶抑制剂(ACEI)、他汀类药物、糖皮质激素、细胞毒性药物、免疫抑制剂。

1.2.2 血液检验指标

高密度脂蛋白胆固醇(HDL-C)、低密度脂蛋白胆固醇(LDL-C)、eGFR、血尿酸(UA)、ALB、载脂蛋白 A1(apoA1)、载脂蛋白 B(apoB)、脂蛋白 a[Lp(a)]、纤维蛋白原(Fib)。血液检测仪器为贝克曼库尔特 AU5800 系列全自动生化分析仪(分光光度测定法和电势测定法)，检测数值取心血病管患者出现结局前，对照组取 5 年随访期间检测记录平均水平。

1.2.3 观察及评价指标

主要对随机森林模型相关的以下参数和指标进行观察与评价：(1)随机森林模型参数，节点值(mtry)、决策树数目(ntree)；(2)变量的预测重要性指标，Gini 值平均降低量(mean decrease gini)；(3)模型预测性能评估，准确率(accuracy)＝(真阳性＋真阴性)/(全部测试集)×100%；精确率(precision)＝真阳性/(真阳性＋假阳性)×100%；召回率(recall)＝真阳性/(真阳性＋假阴性)×100%；ROC 曲线下的面积(AUC)。

1.3 统计学处理

研究数据分析应用 R 软件(版本 4.1.1)进行。其中，呈正态分布的连续资料以 $\bar{x}\pm s$ 表示，非正态的连续资料以中位数[四分位间距]表示，分类资料以例(百分比)表示。视变量类型应用 Wilcoxon 秩和检验、卡方检验比较基线特征。通过训练集数据进行随机森林预测模型构建，选取最优 mtry 和 ntree，旨在降低模型预测错误率。用测试集数据进行模型验证，观察变量的预测重要性并评价模型预测性能。以双侧的 $P<0.05$ 为差异有统计学意义。

2 结 果

2.1 基线资料比较

本研究中训练集 251 例，测试集 99 例，观察结局患心血管病者分别为 80 例(占 31.9%)、35 例(占 35.4%)，所占比例不代表心血管疾病发病率。两组间 BMI、apoB、细胞毒性药物使用存在差异，其余指标均未见明显差异，见表 1。

表 1 NS 患者训练集与测试集基线信息对比

变量	训练集(n=251)	测试集(n=99)	P
一般情况			
有心血管疾病[n(%)]	80 (31.9)	35 (35.4)	0.532
男[n(%)]	159 (63.3)	54 (54.5)	0.129

续表 1　NS 患者训练集与测试集基线信息对比

变量	训练集(n=251)	测试集(n=99)	P
年龄($\bar{x}\pm s$,岁)	49.47±20.36	51.52±19.19	0.380
BMI[$M(P_{25},P_{75})$,kg/m^2]	24.6(23.0,26.9)	24.2(21.9,26.0)	0.005
少数民族[n(%)]	26(10.4)	16(16.2)	0.132
吸烟[n(%)]	74(29.5)	31(31.3)	0.736
饮酒[n(%)]	71(28.3)	30(30.3)	0.708
既往病史[n(%)]			
外周动脉粥样硬化	25(10.0)	10(10.1)	0.968
血糖升高[a]	33(13.1)	12(12.1)	0.796
高血压	108(43.0)	37(37.4)	0.997
静脉血栓疾病	11(4.4)	10(10.1)	0.042
药物使用情况[n(%)]			
抗血小板药	114(45.4)	53(53.5)	0.171
抗凝药	71(28.3)	26(26.3)	0.703
人血白蛋白	103(41.0)	48(48.5)	0.205
ARB	180(71.7)	67(67.7)	0.455
ACEI	43(17.1)	12(12.1)	0.246
他汀类	192(76.5)	80(80.8)	0.382
糖皮质激素	205(81.7)	88(88.9)	0.100
细胞毒性药物	61(24.3)	43(43.4)	0.001
免疫抑制剂	69(27.5)	33(33.3)	0.279
血液检验指标[$M(P_{25},P_{75})$]			
HDL-C(mmol/L)	2.06(1.50,2.54)	1.90(1.44,2.30)	0.122
LDL-C(mmol/L)	4.23(3.23,5.34)	3.96(3.06,4.58)	0.102
eGFR(mL·min^{-1}·1.73 m^{-2})	86.09(64.17,101.89)	78.05(61.95,97.28)	0.150
UA(μmol/L)	525.00(444.50,642.00)	522.00(431.50,653.00)	0.992
ALB(g/L)	42.80(33.75,48.55)	43.30(32.80,47.80)	0.976
apoA1(g/L)	1.83(1.47,2.16)	1.64(1.31,2.06)	0.100
apoB(g/L)	1.87(1.43,2.46)	1.72(1.23,2.18)	0.044
Lp(a)(mg/L)	586.00(303.40,1036.45)	700.00(342.75,986.65)	0.670
Fib(g/L)	4.88(4.03,6.08)	4.64(3.86,5.64)	0.148

[a]：血糖升高包括糖尿病和药物性高血糖(使用激素者)。

2.2　随机森林模型评价指标

随机森林模型最佳 mtry 为 6、ntree 为 446，取该参数时模型错误率最低(图 1)。本研究尝试使用 Gini 值平均降低量作为随机森林模型中变量重要性的衡量标准(图 2)，进一步确定 NS 患者发生心血管疾病结局的重要预测因子。本模型中的相对重要预测因子依此为：eGFR、年龄、HDL-C、apoB、ALB、apoA1、Fib、UA、LDL-C，变量 Gini 值平均降低量与其在模型中的重要性呈正比。本研究构建的预测模型的准确率为 0.919、精确率为 0.935、召回率为 0.829。绘制模型 ROC 曲线(图 3)，AUC 及 95%CI 为 0.899(0.832～0.966)。

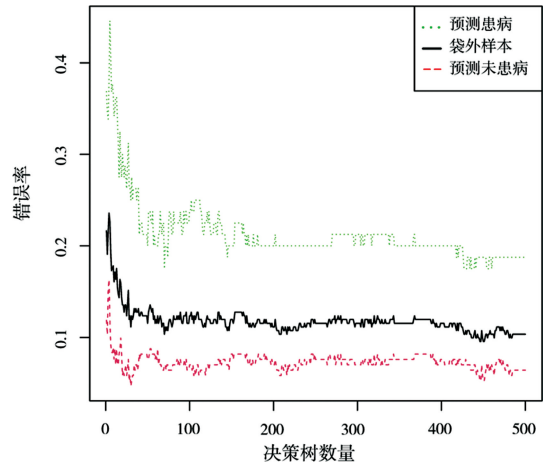

图 1　模型错误率与决策树数量的关系图

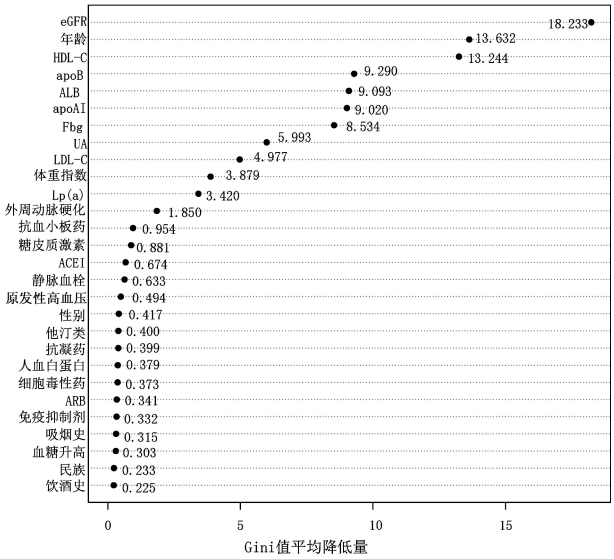

图 2 变量预测重要性示意图

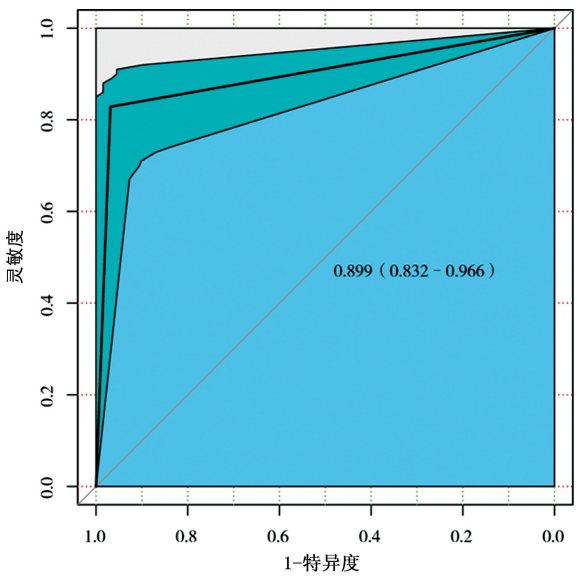

图 3 随机森林模型 ROC 曲线和置信区间图

3 讨 论

NS 有不同程度的甘油三酯、胆固醇和载脂蛋白升高等脂质代谢紊乱，导致动脉内膜脂质浸润，增加了动脉粥样硬化的风险，是 NS 并发心血管事件的危险因素[4]。尤其在频繁复发型或类固醇耐药型 NS 患者中，可能因长期暴露于高脂血症、高氧化应激、频繁感染、持续蛋白尿、低清蛋白血症、血栓栓塞、类固醇、非甾体类药物和免疫抑制剂的不良反应（脂代谢紊乱、血管毒性的和肾毒性）等多种风险因素下，进而引发血管内皮功能受损甚至增加心血管不良事件风险[5]。在当前医疗和研究背景下，仍然没有系统合理的 NS 患者心血管风险预测模型相关研究。即便借用慢性肾病心血管风险预测模型，但随着诊疗技术的发展，传统的心血管危险因素在预测临床结果方面的作用减弱，多数模型在慢性肾病患者中校准不佳，且直接应用于 NS 患者亦存在不合理性[6]。亟须一种预测模型算法，可以处理大量真实世界中错综复杂的预测因子，以达到精准预测且方便获取临床信息的目的。当前随机森林算法正广泛应用于具有大量预测因子数据集的医学预测模型开发，其优秀的数据处理能力和预测性能得到越来越多学者的认可[7]。本研究应用机器学习算法中的随机森林模型对 350 例 NS 患者进行 5 年的心血管疾病风险预测，模型纳入 28 个临床上易获取的预测变量，验证得 ROC 为 0.899 展现出优秀的预测性能，模型召回率 0.829，提示模型对正例的识别能力良好。

本模型中的相对重要预测因子与传统心血管风险因素互有异同。血浆脂质一直以来是心血管风险研究最常用也最易获得的预测因子[8]。NS 中的脂质异常主要是由于脂质清除受损，而不是由于生物合成增加[9]。包括血浆胆固醇、甘油三酯、脂蛋白[乳糜微粒（CM）、极低密度脂蛋白（VLDL）、LDL、中间密度脂蛋白（IDL）和 Lp(a)]水平升高。HDL-C 水平正常或降低，载脂蛋白 apoA1、apoB、apoC 和 apoE 等水平升高[9]。研究表明，在他汀类药物治疗的患者中，apoB 是比 LDL-C 更准确的心肌梗死风险标志物[10]。在本研究中 apoB 在所有脂质中所占重要性也排在首位，提示临床医生在监测患者血脂动态变化时不应忽略这项指标。观察性研究已反复证明 HDL-C 水平与心血管预后之间存在负相关[11]。ApoA1 是 HDL 中含量最丰富的蛋白质，它调节影响 HDL 的心脏保护功能的相互作用[12]。既往研究支持 HDL-C、apoA1 在本研究模型中占有较高重要性的发现。长期以来，LDL-C 都被认为心血管风险因素中最重要的脂质，也是主要的可改变因素。最近欧洲和美国的多社会血脂异常指南强调了降低 LDL-C 对降低心血管风险的重要性[13]。虽然在本模型中 LDL-C 重要性不高，但作为模型预测因子，LDL-C 是一项不可忽略的变量。

除脂质谱外，本研究同样发现其他血液监测指标在模型预测中的重要性。既往研究表明，随着 eGFR 降至 60.00～75.00 mL·min^{-1}·1.73 m^{-2} 以下，发生冠心病的概率呈线性增加[14]，但 eGFR 并未被正式纳入肾脏特异性预测变量[14]。本研究中训练集 eGFR 中位值为 86.09 mL·min^{-1}·1.73 m^{-2}，提示即便 eGFR 未下降至 60.00～75.00 mL·min^{-1}·1.73 m^{-2} 以下亦可以作为 NS 患者预测心血管风险最重要的指标（Gini 值平均降低量 18.233），进一步可能需要基于更大样本的预测模型验证本研究的观点。ALB 是一种有用的心血管疾病风险分层工具，包括急性冠状动脉综合征或心力衰竭，且与稳定性冠心病患者心血管事件发生率呈正相关[15]。Fib 是一种已知的心血管疾病风险标志物，不仅与心血管病状态相关，而且还有助于预测随访时的全因和心血管死亡率[16]。尿酸升高与传统心血管风险、代谢综合征、胰岛素抵抗和慢性肾脏疾病有关[17]。上述观点均提示，本研究构建的随机森林模型中重要预测因子是合理的。

本研究中他汀类药物使用在预测变量中所占的重要性不高，并不能说明他汀类药物对心血管风险方面无显著影响，可能因大多数患者脂质谱管理不佳，

诸多指标掩盖了统计分析中他汀类对结局的影响。同样，本研究中使用各种药物（如抗血小板药、抗凝药、类固醇、免疫抑制剂、细胞毒性药物、ACEI、ARB等）在随机森林预测模型中所占重要性不高，考虑可能为应用药物者占总样本比例偏高，鉴于当前 NS 患者临床诊疗不断规范化，药物使用情况或许并不影响随机森林模型的构建。

为不遗漏对心血管病结局的预测，作者认为召回率是评价该模型预测性能更好的指标。尽管所提出的模型在临床实际使用前需要进一步改进，但随机森林分类算法确定的重要预测因子可能为预测 NS 患者 5 年心血管风险提供有用的信息，可根据本研究筛选出的重要因子进一步开发临床预测模型。应用模型预测患者心血管疾病风险并及时合理地进行干预，为随访期间检测指标的选择提供依据，对于合理利用医疗资源、改善患者预后具有重要意义。

参考文献

[1] WANG C S, GREENBAUM L A. Nephrotic syndrome[J]. Am Fam Physician, 2016, 93(6): Online.

[2] GO A S, TAN T C, CHERTOW G M, et al. Primary nephrotic syndrome and risks of eskd, cardiovascular events, and death: the kaiser permanente nephrotic syndrome study[J]. J Am Soc Nephrol, 2021, 32(9): 2303-2314.

[3] LEE T, DEREBAIL V K, KSHIRSAGAR A V, et al. Patients with primary membranous nephropathy are at high risk of cardiovascular events[J]. Kidney Int, 2016, 89(5): 1111-1118.

[4] FERENCE B A, GINSBERG H N, GRAHAM I, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel[J]. Eur Heart J, 2017, 38(32): 2459-2472.

[5] AGRAWAL S, ZARITSKY J J, FORNONI A, et al. Dyslipidaemia in nephrotic syndrome: mechanisms and treatment[J]. Nat Rev Nephrol, 2017, 14(1): 70.

[6] STREJA E, NORRIS K C, BUDOFF M J, et al. The quest for cardiovascular disease risk prediction models in patients with nondialysis chronic kidney disease[J]. Curr Opin Nephrol Hypertens, 2021, 30(1): 38-46.

[7] SPEISER J L. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data[J]. J Biomed Inform, 2021, 117: 103763.

[8] SAEED A, FEOFANOVA E V, YU B, et al. Remnant-Like particle cholesterol, Low-Density lipoprotein triglycerides, and incident cardiovascular disease[J]. J Am Coll Cardiol, 2018, 72(2): 156-169.

[9] HARI P, KHANDELWAL P, SMOYER W E. Dyslipidemia and cardiovascular health in childhood nephrotic syndrome[J]. Pediatr Nephrol, 2020, 35(9): 1601-1619.

[10] JOHANNESEN C, MORTENSEN M B, LANGSTED A, et al. Apolipoprotein B and Non-HDL cholesterol better reflect residual risk than LDL cholesterol in Statin-Treated patients[J]. J Am Coll Cardiol, 2021, 77(11): 1439-1450.

[11] SMIT R A, JUKEMA J W, TROMPET S. Increasing HDL-C levels with medication: current perspectives[J]. Curr Opin Lipidol, 2017, 28(4): 361-366.

[12] COOKE A L, MORRIS J, MELCHIOR J T, et al. A thumbwheel mechanism for APOA1 activation of LCAT activity in HDL[J]. J Lipid Res, 2018, 59(7): 1244-1255.

[13] ATAR D, JUKEMA J W, MOLEMANS B, et al. New cardiovascular prevention guidelines: How to optimally manage dyslipidaemia and cardiovascular risk in 2021 in patients needing secondary prevention?[J]. Atherosclerosis, 2021, 319: 51-61.

[14] SARNAK M J, AMANN K, BANGALORE S, et al. Chronic kidney disease and coronary artery disease: JACC State-of-the-Art review[J]. J Am Coll Cardiol, 2019, 74(14): 1823-1838.

[15] SUZUKI S, HASHIZUME N, KANZAKI Y, et al. Prognostic significance of serum albumin in patients with stable coronary artery disease treated by percutaneous coronary intervention[J]. PLoS One, 2019, 14(7): e0219044.

[16] PIETERS M, FERREIRA M, DE MAAT M, et al. Biomarker association with cardiovascular disease and mortality - The role of fibrinogen. A report from the NHANES study[J]. Thromb Res, 2021, 198: 182-189.

[17] NDREPEPA G. Uric acid and cardiovascular disease[J]. Clin Chim Acta, 2018, 484: 150-163.

（收稿日期：2021-04-21 修回日期：2021-09-21）