

人工智能自然语言处理诊断肾癌 T 分期的研究

温 力 江卫星 孙丰龙 毕新刚 寿建忠 马建辉

摘 要 **目的** 探讨人工智能自然语言处理方法诊断肾癌 T 分期的可行性和准确性。**方法** 收集 2018 年 1 月~2020 年 1 月国家癌症中心、国家肿瘤临床医学研究中心、中国医学科学院/北京协和医学院肿瘤医院 200 例肾癌患者作为训练组,选取 2015 年 1 月~2017 年 12 月性别、年龄、病理分期匹配的 200 例患者作为测试组,使用基于规则匹配和条件随机场两种人工智能自然语言处理方法对病理文本数据进行提取分析。**结果** 基于规则匹配和条件随机场两种方法在测试组的正确率分别为 99.0% 和 95.5%。测试组的方法性能评估中,规则匹配方法的准确率为 99.0%,召回率为 99.0%, F_1 -分数为 99.0%。条件随机场方法的准确率为 97.1%,召回率为 95.5%, F_1 -分数为 96.3%。**结论** 人工智能通过自然语言处理方法自动诊断肾癌 T 分期可行,且基于规则匹配的算法准确性较高,这些研究结果有待于多中心、大样本数据进一步验证。

关键词 人工智能 诊断 肾癌 分期

中图分类号 R73

文献标识码 A

DOI 10.11969/j.issn.1673-548X.2023.04.026

Research on Artificial Intelligence Natural Language Processing in Diagnosing T-stage of Renal Carcinoma. WEN Li, JIANG Weixing, SUN Fenglong, et al. National Cancer Center/National Clinical Research Center For Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

Abstract **Objective** To explore the feasibility and accuracy of artificial intelligence natural language processing in diagnosing T-stage of renal carcinoma. **Methods** A total of 200 patients with renal carcinoma from National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College from January 2018 to January 2020 were selected as the training group, and 200 patients with matched gender, age and pathological stage between January 2015 to December 2017 were selected as the test group. Two artificial intelligence natural language processing methods including rule based template matching and conditional random fields were used to extract and analyze pathological text data. **Results** The accuracy rates of rule-based template matching and conditional random fields were 99.0% and 95.5% respectively. In the method performance evaluation of the test group, the accuracy rate of the rule-based template matching was 99.0%, the recall rate was 99.0%, and the F_1 -score was 99.0%. The accuracy rate of the conditional random fields was 97.1%, the recall rate was 95.5%, and an F_1 -score was 96.3%. **Conclusion** It is feasible for artificial intelligence to automatically diagnose the T-stage of renal carcinoma by natural language processing, and the algorithm of rule-based template matching had high accuracy. These research results still need to be further verified by multi-center and large sample data.

Key words Artificial intelligence; Diagnosis; Renal carcinoma; Stage

人工智能是模拟人类智能的一种算法过程,随着科技的进步和发展,人工智能在医学中的研究正蓬勃发展^[1,2]。自然语言处理(natural language processing, NLP)作为人工智能的一个分支领域,是将语言转换为计算表达的形式,可作为工具用于处理肿瘤诊疗信息,特别是在提取临床电子病历(electronic medical records, EMR)文本数据中发挥了重要作用^[3]。

基于 NLP 构建自由文本挖掘算法去进一步挖掘肿瘤临床信息是十分有前景的新方法,比如提取肿瘤的分期、分级等数据^[4]。在临床工作中,恶性肿瘤的分期是预后评估和制定治疗计划的重要程序之一,美国癌症联合委员会(American Joint Committee on Cancer, AJCC)制定的 TNM 分期是最常用的癌症分期系统。肾癌作为泌尿系统常见的恶性肿瘤,其分期依据主要包括原发肿瘤直径、肾周脂肪、肾盂、静脉瘤栓、区域淋巴结侵犯以及远处转移。这些临床资料大部分能从病理报告中获得,而病理报告是传达肿瘤分期的黄金标准,可以以自由文本形式存储,有研究者基于 NLP 建立了肺癌英文的 T 分期算法^[5]。本研究利用

作者单位:100021 国家癌症中心、国家肿瘤临床医学研究中心、中国医学科学院/北京协和医学院肿瘤医院(温力、毕新刚、寿建忠、马建辉);100051 首都医科大学附属北京友谊医院(江卫星);100080 北京,神州医疗科技股份有限公司(孙丰龙)

通信作者:毕新刚,电子邮箱:bixingang@cscsco.org.cn

NLP 作为工具对笔者医院收治的肾癌患者临床电子病历数据进行人工智能自然语言处理分析,提出一种可自动诊断肾癌 T 分期的算法,以进一步协助病理医生进行病理诊断。

资料与方法

1. 临床资料收集:收集 2018 年 1 月~2020 年 1 月因肾癌于国家癌症中心、国家肿瘤临床医学研究中心、中国医学科学院/北京协和医学院肿瘤医院行手术治疗的 200 例肾癌患者作为训练组,并选取 2015 年 1 月~2017 年 12 月性别、年龄以及病理分期匹配的 200 例因肾癌行手术患者作为测试组。病理结果均以自由文本报告的形式存储在临床诊疗信息系统中,每份报告由几个结构部分组成,包括肿瘤切除部位、性质、分级、肿瘤直径、侵犯范围等。所有患者均有完整的病理诊断和病理分期。

2. 统计学方法:临床及病理一般数据应用 SPSS 25.0 统计软件分析,计量资料符合正态分布,以均数 $\pm$ 标准差($\bar{x} \pm s$)表示,不符合正态分布以中位数(范围)[M(min~max)]表示。计数资料以例数(百分比)[n(%)]表示,组间比较采用 χ^2 检验。肿瘤的病理分期参考 2017 年第 8 版 TNM 分期标准^[6]。将病理医生复阅病理报告的病理分期定义为金标准。使用基于规则模板匹配和基于条件随机场的两种 NLP 算法,采用 python=3.6 和 sklearn_crfsuite=0.3.6 进行开发信息抽取算法,并对两种方法的预测效果进行对比。所有数据先进行预处理步骤使数据适合分析,然后进行处理步骤,处理步骤为实际算法。

3. 概念框架:研究框架设计如图 1 所示,将数字化肾癌病理报告作为输入,针对肾癌 T 分期组成的各

个要素,病理医生使用训练组为各分期要素创建一组同义词及其词缀,即表 1 中的关键词,同时将分期要素定义为识别域及其值,如“肾周脂肪”为识别域,则“累及”和“未累及”为值。开发基于规则模板匹配和基于条件随机场的两种信息抽取算法。通过图 2 所示流程依据抽取的信息获得肾癌病理 T 分期的预测结果,并且在测试组数据上计算预测准确性,作为分期预测模型性能的整体评估。

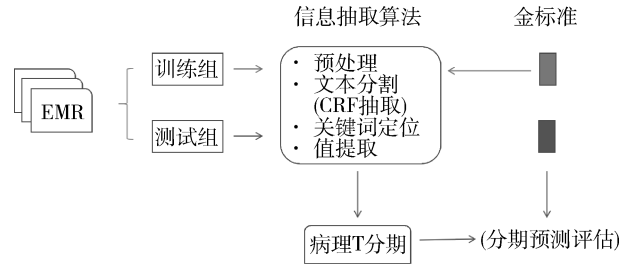


图 1 基于 NLP 的中文病理报告处理框架及在病理 T 分期预测中的应用

表 1 肾癌病理报告中识别域对应的关键词和值提取规则

识别域	关键词	规则
肿瘤直径	肿瘤 肿瘤大小 最大径	R ₁
肾上腺	肾上腺	R ₂ , R ₃
肾筋膜外	腰大肌 肠壁 肠管 腹膜 脾脏 胰腺 椎体 肝脏	R ₂ , R ₃
腔静脉	腔静脉	R ₂ , R ₃
腔静脉壁	腔静脉壁	R ₂ , R ₃
肾盂	肾盂	R ₂ , R ₃
肾静脉	肾静脉 肾门血管 肾段静脉	R ₂ , R ₃
肾周脂肪	肾周脂肪 周围脂肪 肾脂肪囊	R ₂ , R ₃
肾窦脂肪	肾窦脂肪	R ₂ , R ₃

“|”分隔多个可选项。针对值的提取总体规则为 R₂, R₃, 不同部位的具体规则略有差别。如肾静脉 R₃ 中“特定描述”为“瘤栓”,而肾盂 R₃ 中的“特定描述”一般为“累及”和“侵犯”

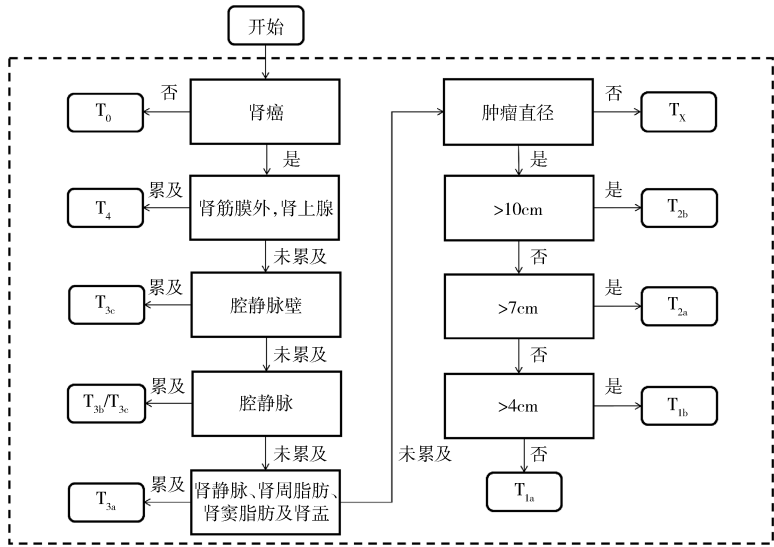


图 2 基于识别域值提取的肾癌病理 T 分期预测流程图

4. 基于规则模板匹配的方法:算法分为4个部分,包括预处理、文本分割、关键词识别和值提取。首先,在预处理中统一标点符号,将英文的逗号、分号、句号和括号转化为相应的中文符号。然后,根据中文书写习惯,以逗号、分号和句号作为分界符号,将预处理后

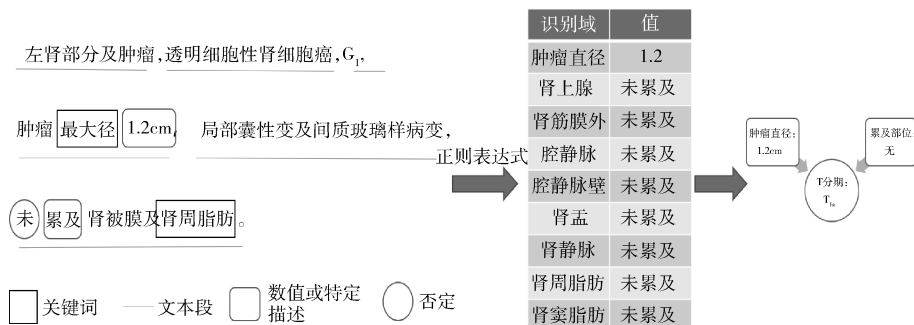


图3 肾癌病理报告病理T分期关键信息抽取示例

具体建立规则模板的过程如下:针对提取出来的包含关键词的文本小片段,定位一些规律性的表达,例如,肿瘤直径的文字描述一般是关键词“肿瘤大小”,“最大径”或者“肿瘤”后面加上相应的数值表达式;肿瘤累及部位的文字描述一般为具体部位前面或者后面加上特定的描述词如“累及”、“侵犯”、“见瘤”、“瘤栓”等;而关于肿瘤未累及部位的文字描述,则是在特定描述上再添加否定词如“未”。笔者研究使用python正则表达式设计以下提供进行匹配的模板:(R₁)关键词+数值,例如:最大径2.5cm;(R₂)特定描述+关键词,例如:侵犯肾上腺;(R₃)关键词+特定描述,例如:肾静脉瘤栓。对肿瘤累及部位的描述中如出现否定词则排除,并在正则表达式中体现。

5. 基于条件随机场的方法:对照着基于规则模板匹配的方法,用条件随机场替代文本分割部分和值提取中的语义识别部分。先是预处理统一标点符号,然后使用开源软件brat标注出病理报告中对于判断肾癌病理T分期有用的陈述信息:(1)大小陈述,如“肿瘤最大径6.5cm”。(2)累及陈述,如“肾门处可见血管侵犯,形成瘤栓”。(3)未累及陈述,如“未累及肾周脂肪、肾窦脂肪及肾盂”,详见图4。类似于命名实体识别,再使用“BMEIO”用于定义每个有用陈述信息的边界。其中“B”代表起始字符,“E”代表结束字符,“M”是中间字符,“O”代表其他无关字符。同时,在边界信息后面加上陈述类型,所以一个字符完整的标注形如“B-大小陈述”或者“O”。模型在预测的时候,将预测为同一类陈述的“B”,紧随其后的一系列“M”以及之后的“E”的字符进行连续组合即为所

的文本分割为一些小片段。接着,针对每个关键词,通过字符串匹配定位出与它相关的文本小片段。最后,对训练组数据中定位出来的文本小片段总结出规则模板,设计正则表达式进行规则匹配,提取每个识别域对应的值。信息抽取及T分期预测过程示例如图3所示。

提取的陈述。

在对训练组完成标注后,研究使用sklearn_crfsuite=0.3.6中的条件随机场进行训练得到模型,用模型对测试组中数据进行预测,提取出有用陈述。在条件随机场提取出有用陈述之后,依据陈述的类别信息,通过简单的数字提取和关键词搜索就可以得到肿瘤直径及累及部位信息,从而完成值提取过程。

左肾及肿瘤,肾透明细胞性肾细胞癌,fuhrman分级,II级。

大小陈述 累及及陈述
肿瘤最大径6.5cm, 累及肾被膜, 未累及肾周脂肪、肾窦脂肪及肾盂,
累及陈述
肾门处可见血管侵犯,形成瘤栓,未见明确神经侵犯。
未见明确神经侵犯。输尿管切缘未见肿瘤。周围肾未见异常。

图4 条件随机场方法标注示例

6. 肾癌T分期预测效能评估:在测试组上,基于不同的信息抽取值算法及流程图2可以得到病理T分期的预测值,通过与金标准比较,计算整体的准确率(precision)、召回率(recall)、F₁-分数(F₁-score)以及正确率(accuracy)^[7]。其中正确率,即预测正确的数目与全部病例数目之比。对于准确率、召回率和F₁-分数,设FP为假阳数目,TP为真阳数目,FN为假阴数目,TN为真阴数目,则

$$\text{Precision} = \frac{TP}{TP + FP};$$

$$\text{Recall} = \frac{TP}{TP + FN};$$

$$F_1 - \text{score} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

因为研究的问题为多分类问题,而且类别不平衡,所以针对每一个类别计算出对应的准确率(记为 Precision_i)、召回率(记为 Recall_i)和 F_1 -分数(记为 $F_1\text{-score}_i$)后,再根据各类样本比例配置权重得到加权平均值,作为总的准确率、召回率和 F_1 -分数,即

$$\begin{aligned}\text{Precision} &= \sum_i w_i \text{Precision}_i; \\ \text{Recall} &= \sum_i w_i \text{Recall}_i; \\ F_1\text{-score} &= \sum_i w_i F_1\text{-score}_i.\end{aligned}$$

结 果

1. 肾癌手术患者的临床病理特征:本研究共 400 例患者,其中男性 287 例,女性 113 例;患者年龄 25 ~ 80 岁,中位年龄 62 岁;肿瘤直径为 $6.1 \pm 2.3\text{cm}$ 。透明细胞癌 356 例,乳头状细胞癌 22 例,嫌色细胞癌 13 例,其他病理类型包括 MIT 家族转位性肾癌、集合管癌等共 9 例。根据 TNM 分期标准, pT_{1a} 期 252 例, pT_{1b} 期 69 例, pT_{2a} 期 10 例, pT_{2b} 期 4 例, pT_{3a} 期 59 例, pT_{3b} 期 2 例, pT_4 期 4 例(表 2)。

表 2 400 例患者一般临床资料 [$n(\%)$, $\bar{x} \pm s$, $M(\text{min} \sim \text{max})$]

临床病理特征	数值
性别	
男性	287(71.8)
女性	113(28.2)
年龄(岁)	62(25 ~ 80)
肿瘤位置	
左	198(49.5)
右	202(50.5)
肿瘤直径(cm)	6.1 ± 2.3
病理类型	
透明细胞癌	356(89.0)
乳头状细胞癌	22(5.5)
嫌色细胞癌	13(3.3)
其他	9(2.2)
病理分期	
T_{1a}	252(63.0)
T_{1b}	69(17.3)
T_{2a}	10(2.5)
T_{2b}	4(1.0)
T_{3a}	59(14.7)
T_{3b}	2(0.5)
T_{3c}	0(0)
T_4	4(1.0)

2. 测试组结果分析:根据测试组 200 例患者效能评估结果,基于规则模板匹配方法准确率为 99.0%,召回率为 99.0%, F_1 -分数为 99.0%,正确率为 99.0%;而基于条件随机场方法,准确率为 97.1%,召回率为 95.5%, F_1 -分数为 96.3%,正确率为 95.5%(表 3)。前者正确率高于后者(99.0% vs

95.5%, $\chi^2 = 4.581$, $P = 0.032$)。

表 3 肾癌病理报告 T 分期预测结果(%)

项目	准确率	召回率	F_1 -分数	正确率
基于规则模板匹配	99.0	99.0	99.0	99.0
基于条件随机场	97.1	95.5	96.3	95.5

讨 论

笔者之前总结了国内外人工智能诊断肾细胞癌的现状,关于人工智能在肾细胞癌中的研究和应用目前尚处于起步阶段,主要集中在影像学,而病理研究较少报道^[8]。本研究对肾癌病理报告利用人工智能自然语言处理方法进行识别和分析,提出可自动诊断肾癌 T 分期的算法,从而降低病理医生负荷,提高临床工作效率。

近年来,越来越多的研究证实人工智能在医学研究中展现出了巨大的优势,特别是在大数据的趋势下,人工智能作为一种自动化的工具,在多个方面直接或者间接地减轻医疗工作的负担^[9~11]。通过医学图片的人工智能深度学习可协助诊断,如 Campanella 等^[12]开发了多层面的人工智能决策支持系统,对共 15187 例癌症患者包括 44732 张病理切片进行了测试,诊断准确性的曲线下面积达 0.98,预计能减少 70% 病理医生的工作量。通过病历文本的自然语言处理可协助肿瘤分期,如 Nguyen 等^[5]使用基于规则的符号分类从 718 例肺癌的自由文本病理报告中实现对肺癌自动分期,准确性可达到 80% 左右。在泌尿肿瘤方面, Schroeck 等^[13]报道了从 600 份膀胱癌病理报告中提取信息的基于规则方法的准确性在 68% ~ 98%。 Odisho 等^[14]开发了一种对前列腺癌病理报告和临床指标的识别算法,准确性为 94.8% ~ 100%。而缺乏肾癌的自然语言文本处理的相关研究报道。

本研究利用自然语言方法对肾癌的病理报告进行了提取分析,首先收集 2018 年 1 月 ~ 2020 年 1 月 200 例肾癌病理报告作为训练组,既往 2015 年 1 月 ~ 2017 年 12 月年龄、性别、病理分期匹配的 200 例肾癌病理报告作为测试组,然后分别使用基于规则模板匹配和基于条件随机场两种方法预测肾癌的 T 分期,结果显示,两者在测试组的正确率分别为 99.0% 和 95.5% ($P = 0.032$)。在测试组的方法性能评估中,对于规则匹配方法的准确度为 99.0%,召回率为 99.0%, F_1 -分数为 99.0%;对于条件随机场方法的准确度 97.1%,召回率为 95.5%, F_1 -分数为 96.3%。从以上结果的准确性来看,两种不同的自然语言处理

方法均是可行的,使用规则匹配的方法可能略好。但笔者发现规则匹配的方法虽然实现了关键信息的提取,如作为处理更大的数据集或更换癌种后,可能需要设计更复杂的规则模板,而基于条件的随机场法可自主学习标注陈述的关键信息,可能在大数据上具有较好的适应性优势。

笔者对在本研究中预测错误的病理报告也进行了分析,错误的识别主要源自病理报告内容不完整,在按照规则流程进行分期时,因缺乏识别域和值导致识别错误,如病理报告中未描述肿瘤直径,则识别为 T_x 。因此,笔者认为完整的病理报告是提高准确程度的必要因素,这样也有利于规范病理医生的报告书写,根据病理报告识别内容可形成病理报告书写统一规范,适合在基层推广,提高基层医疗水平。另外,根据本研究结果可开发相应的软件系统,协助病理医生进行分期,大大减少病理医生的工作量。

本研究存在一定的局限性:(1)研究来自单中心的病例,病理报告书写规范比较统一,仍需与其他机构合作进行完善识别工作。(2)本研究入组病例数量有限,准确度可能有偏差,需进一步收集大样本量的病理资料去验证。(3)本研究仅为单个癌种的识别算法,对其他癌种并不适用,笔者将在下一步利用本体语言去扩大适用范围。

综上所述,使用自然语言处理通过病理报告诊断肾癌病理 T 分期是可行的方法,基于规则匹配的算法具有较高准确度。但这些研究结果有待下一步多中心、大样本量的数据去证实。国内外尚无肾癌病理报告人工智能识别自动诊断 T 分期的报道,本研究可作

为未来研究的初步参考。

参考文献

- Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists[J]. JAMA, 2016, 316(22): 2353 – 2354
- Topol EJ. High – performance medicine: the convergence of human and artificial intelligence[J]. Nat Med, 2019, 25(1): 44 – 56
- Savova GK, Danciu I, Alamudun F, *et al.* Use of natural language processing to extract clinical cancer phenotypes from electronic medical records[J]. Cancer Research, 2019, 79(21): 5463 – 5470
- Nobel JM, Puts S, Bakers FCH, *et al.* Natural language processing in dutch free text radiology reports: challenges in a small language area staging pulmonary oncology[J]. J Digit Imaging, 2020, 33(4): 1002 – 1008
- Nguyen AN, Lawley MJ, Hansen DP, *et al.* Symbolic rule – based classification of lung cancer stages from free – text pathology reports[J]. J Am Med Inform Assoc, 2010, 17(4): 440 – 445
- Shao N, Wang HK, Zhu Y, *et al.* Modification of American Joint Committee on cancer prognostic groups for renal cell carcinoma[J]. Cancer Medicine, 2018, 7(11): 5431 – 5438
- Wang H, Zhang W, Zeng Q, *et al.* Extracting important information from Chinese Operation Notes with natural language processing methods[J]. J Biomed Inform, 2014, 48: 130 – 136
- 江卫星, 郑闪, 寿建忠, 等. 人工智能在肾细胞癌诊断中的研究现状[J]. 中华泌尿外科杂志, 2020, 41(3): 233 – 236
- Bera K, Schalper KA, Rimm DL, *et al.* Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology[J]. Nat Rev Clin Oncol, 2019, 16(11): 703 – 715
- Wallis C. How artificial intelligence will change medicine[J]. Nature, 2019, 576(7787): S48
- Couzin – Frankel J. Medicine contends with how to use artificial intelligence[J]. Science, 2019, 364(6446): 1119 – 1120
- Campanella G, Hanna MG, Geneslaw L, *et al.* Clinical – grade computational pathology using weakly supervised deep learning on whole slide images[J]. Nature Medicine, 2019, 25(8): 1301 – 1309
- Schroek FR, Patterson OV, Alba PR, *et al.* Development of a natural language processing engine to generate bladder cancer pathology data for health services research[J]. Urology, 2017, 110: 84 – 91
- Odisho AY, Bridge M, Webb M, *et al.* Automating the capture of structured pathology data for prostate cancer clinical care and research[J]. JCO CCI, 2019, 3: 1 – 8
- (收稿日期: 2022 – 07 – 31)
- (修回日期: 2022 – 08 – 07)
- Oh J, Ahn J. Comparison of retinal layer thickness and vascular density between acute and chronic branch retinal vein occlusion[J]. Korean J Ophthalmol, 2019, 33(3): 238 – 248
- 王林妮, 于荣国, 李筱荣, 等. 视网膜分支静脉阻塞黄斑水肿程度对黄斑中心凹无血管区量化分析的影响[J]. 中华眼底病杂志, 2019, 35(1): 20 – 24
- Parodi MB, Visintin F, Della Rupe P, *et al.* Foveal avascular zone in macular branch retinal vein occlusion[J]. Int Ophthalmol, 1995, 19(1): 25 – 28
- Kuehlewein L, Tepelus TC, An L, *et al.* Noninvasive visualization and analysis of the human parafoveal capillary network using swept source optical microangiography[J]. Invest Ophthalmol Vis Sci, 2015, 56(6): 3984 – 3988
- Balaratnasingam C, Inoue M, Ahn S, *et al.* Visual acuity is correlated with the area of the foveal avascular zone in diabetic retinopathy and retinal vein occlusion[J]. Ophthalmology, 2016, 123(11): 2352 – 2367
- (收稿日期: 2022 – 04 – 23)
- (修回日期: 2022 – 05 – 02)
- (接第 151 页)
- Wang K, Ghasemi Falavarjani K, Nittala MG, *et al.* Ultra – wide – field fluorescein angiography – guided normalization of ischemic index calculation in eyes with retinal vein occlusion[J]. Invest Ophthalmol Vis Sci, 2018, 59(8): 3278 – 3285
- Huang J, Lu Y, Gu X, *et al.* Correlation between the nonperfusion area on ultra – widefield fluorescein angiography and nonflow area on optical coherence tomographic angiography in retinal vein occlusion[J]. J Ophthalmol, 2021, 2021: 5581319
- Moussa M, Leila M, Bessa AS, *et al.* Grading of macular perfusion in retinal vein occlusion using en – face swept – source optical coherence tomography angiography: a retrospective observational case series[J]. BMC Ophthalmol, 2019, 19(1): 127
- Coscas F, Glacet – Bernard A, Miere A, *et al.* Optical coherence tomography angiography in retinal vein occlusion: evaluation of superficial and deep capillary plexa[J]. Am J Ophthalmol, 2016, 161: 160 – 171
- Chen X, Rahimy E, Sergott RC, *et al.* Spectrum of retinal vascular diseases associated with paracentral acute middle maculopathy[J]. Am J Ophthalmol, 2015, 160(1): 26 – 34, e21