·方法学研究·

面向真实世界的知识挖掘与知识图谱补全研究（三）：基于正则表达式对膀胱癌真实世界数据的结构化信息抽取

马文昊[1,2]，石涵予[3]，黄　桥[1]，黄　兴[4]，王永博[1]，王诗淳[1]，任相颖[1]，施　悦[5]，靳英辉[1]，阎思宇[1]

1. 武汉大学中南医院循证与转化医学中心（武汉 430071）
2. 武汉大学第二临床学院（武汉 430071）
3. 武汉大学弘毅学堂（武汉 430072）
4. 浙江大学医学院附属第一医院泌尿外科（杭州 310003）
5. 武汉大学中南医院信息中心（武汉 430071）

【摘要】随着医疗大数据的发展，真实世界研究近些年来越来越受到重视，发展前景良好，但真实世界研究的实施仍存在一些挑战，引起学者们广泛讨论。真实世界数据的非结构化是目前最亟待解决的问题。本研究以正则表达式为基础，通过基于规则的信息抽取方法对武汉大学中南医院近几年膀胱癌患者的入院记录、病理报告、手术记录和影像记录等数据进行结构化信息抽取，并以准确率和召回率为指标评价其抽取效果，旨在为后续研究提供参考。

【关键词】真实世界数据；信息抽取；正则表达式；自然语言处理；电子病历数据；膀胱癌

Research on real-world knowledge mining and knowledge graph completion (III): structured information extraction from real world data of bladder cancer based on regular expression

MA Wenhao[1,2], SHI Hanyu[3], HUANG Qiao[1], HUANG Xing[4], WANG Yongbo[1], WANG Shichun[1], REN Xiangying[1], SHI Yue[5], JIN Yinghui[1], YAN Siyu[1]

1. Center for Evidence-Based and Translational Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

2. The Second Clinical College of Wuhan University, Wuhan 430071, China

3. HongYi Honor College of Wuhan University, Wuhan 430072, China

4. Department of Urology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China

5. Information Center, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

Corresponding author: JIN Yinghui, Email: jinyinghui0301@163.com; YAN Siyu, Email: 15927071586@163.com

DOI: 10.12173/j.issn.1004-5511.202308006

基金项目：国家自然科学基金面上项目（82174230）；武汉大学中南医院青年交叉学科专项基金（ZNQNTC2023006）

通信作者：阎思宇，实习研究员，助教，Email：15927071586@163.com

【Abstract】With the development of medical big data, the real-world study (RWS) has received increasing attention in recent years, and has a good promising prospect. However, there are still some challenges in the implementation of RWS that has led to extensive discussion among scholars. The most urgent issue currently to be addressed is the unstructured nature of real-world data (RWD). Based on regular expressions, this study used rule-based information extraction method to extract structured information from admission records, pathological reports, surgical records, and image records of bladder cancer patients in Zhongnan Hospital of Wuhan University in recent years, and evaluated the extraction effects with accuracy and recall as indicators, aiming to provide reference for subsequent research.

【Keywords】Real-world data; Information extraction; Regular expression; Natural language processing; Electronic medical record data; Bladder cancer

美国食品和药物监督管理局在《真实世界证据方案的框架》[1]中将真实世界数据（real-world data，RWD）定义为“与患者健康状况有关的和（或）日常医疗过程中收集的各种来源的数据”。RWD 包括来源于卫生信息系统、电子病历（electronic medical record, EMR）、医保系统的数据和来自移动设备端如可穿戴设备获得的相关数据等。随着诊疗数据的几何级增长，基于 EMR 数据开展的真实世界研究越来越受重视，如进行真实环境下干预措施效果和安全性的评价研究[2]，但在实施时仍面临一些挑战。EMR 数据产生的初始目的不是用于临床研究而是服务于临床实践，因此除结构化字段外，还包括大量半结构化、非结构化文本，并且各医疗机构之间数据的记录与储存尚缺乏统一标准，对于数据记录方面的规范化培训和质量控制不足，导致原始数据质量参差不齐，增大了研究者数据挖掘工作的难度。因此如何基于现有 EMR 数据进行结构化信息抽取是一个不小的挑战。

信息抽取作为自然语言处理的子领域，其方法主要包括基于人工编写规则的信息抽取方法和基于统计学方法的信息抽取方法[3]。基于人工编写规则的信息抽取方法相对简单但高度依赖于人工编写的规则集，适用于有一定结构规律的自然语言文本。正则表达式（regular expression, RE）是对字符串操作的一种逻辑公式，即是用事先定义好的一些特定字符及其组合，组成一个“规则字符串”，用以表达对字符串的一种过滤逻辑。RE 是一种文本模式，该模式描述了在搜索文本时要匹配的一个或多个字符串[4]，可以作为一种过滤工具，实现对 RWD 的结构化信息抽取。近些年来，RE 在医学领域有着广泛的应用。例如国外学者应用RE于神经外科手术登记表的构建，显著减少了人工工作量并促进相关临床研究[5]；Flores 等[6]使用 RE 从生物医学文本中提取特征值，有较高的准确性，可为数据集进一步分析奠定基础；在对医学指南中事件句型进行相关匹配与抽取的研究中，RE 可高效准确地将医学指南中的事件自动转换成 XML 结构化数据[7]。

考虑到 EMR 数据中大部分目标字段具有一定的表达规律，故本研究以武汉大学中南医院近 7 年膀胱癌患者 EMR 中的入院记录、病理报告、手术记录和影像记录等非结构化文本数据为例，采用基于人工编写规则并以 RE 为编程基础的信息抽取方法对膀胱癌自然语言文本数据进行结构化信息抽取。

1 资料与方法

1.1 数据源及抽取字段

以武汉大学中南医院 2015—2021 年出院诊断包含膀胱癌的患者 EMR 数据中的入院记录、病理记录、手术记录以及影像学记录为研究数据源，其中病理记录示例见表 1。本研究已通过武汉大学中南医院伦理委员会审核批准（批号：科伦 [2022002K]），所有数据均已进行了去隐私化处理。

在咨询临床、病理、流行病学等专家意见和查阅相关文献后，结合膀胱癌 RWD 的表达规律与结构特点，本研究确定需抽取的结构化字段包括 32 个目标字段（表 2）。其中“肿瘤浸润深度”以膀胱壁的解剖结构层级为标准进行抽取，包括固有层、浅深肌层等；“区域淋巴结浸润情况”

的描述通常涵盖了送检淋巴结位置、送检数量及阳性数量；“变异组织学”采用 2016 年第 4 版《WHO 泌尿系统及男性生殖器官肿瘤分类》[8] 标准进行抽取；当病理记录未明确记录 T、N 分期时，依据美国癌症联合委员会 2017 年制订的第 8 版《膀胱癌 TNM 分期手册》[9]，分别通过对“肿瘤浸润深度”“是否有输尿管残端、输精管断端癌浸润”“是否有精囊腺、前列腺组织、子宫、阴道癌浸润”3 个字段和“区域淋巴结浸润情况”字段的抽取结果推理得到 T、N 分期。“M 分期”也属于推理字段，需要以影像学记录中“肿瘤转移情况”字段抽取结果并结合分期依据推理得到。

1.2 抽取方法

本研究是基于 Python 环境下利用 RE 进行文本信息的抽取。

1.2.1 正则概述

RE 提供对于基本字符、特殊字符、数量词、边界位置等文本的匹配。每一个字符代表不同的匹配规则，例如，字符“\d”表示匹配数字字符；“*”“+”“?”分别表示匹配前一个字符零次或多次、一次或多次、零次或一次；{n, m} 表示匹配至少 n 次，最多 m 次。完整 RE 规则可参考相关文献 [10]。

1.2.2 数据集划分及字段词典编写

使用随机抽样的方法，从入院记录、病理记录、手术记录、影像学记录四个表单中分别抽取 300 条数据的样本，其中 200 条用于规则抽取集，100 条用于评测集。通过人工抽取规则抽取集中的结构化数据，得到目标字段的不同描述，将其归纳总结为规则集，即字段词典（示例见表 3）。以字段词典为规则撰写 RE。

1.2.3 正则抽取实现

在 Python 编译器中，主要通过 Python 强大的库和 re 模块实现正则抽取。本研究整体抽取方法流程图见图 1。

步骤一，数据导入及遍历行信息。使用 pandas 库中的 dataframe，读写 Excel 表格工作簿以及单元格信息，利用 read_excel 对工作簿赋

表1 病理记录示例

Table 1. Example of pathological record

患者号	住院号	病理报告日期	病理诊断
pat_435c102***	ZY01000***	2018/10/29	（膀胱+前列腺+精囊切除标本） 1. 高级别尿路上皮癌，肿瘤侵及膀胱固有肌层，未见明确脉管内瘤栓及神经侵犯 2. 慢性前列腺炎，前列腺增生症 3.（双侧）精囊腺、输精管组织未见癌 4. 送检（左、右输尿管末端）组织未见癌 5. 送检左盆腔（3枚）、右盆腔（2枚）淋巴结未见癌转移 6. 病理分期：pT2N0Mx

表2 抽取的结构化字段

Table 2. Extracted structured fields

文本来源	抽取目标字段
入院记录	“吸烟史：有/无”“吸烟时间（年）”“吸烟史：平均（支/d）”“戒烟：有/无”“戒烟时间（年）”“饮酒史：有/无”“饮酒时间（年）”“饮酒平均：平均（mL/d）”“戒酒史：有/无”“戒酒时间（年）”“毒品接触史”
病理记录	“是否为膀胱标本”“是否为癌症”“是否为膀胱尿路上皮癌”“病理结果中是否记录明确分期”“病理分级”“肿瘤浸润深度”“是否有输尿管残端、输精管断端癌浸润”“是否有精囊腺、前列腺组织、子宫、阴道癌浸润”“基底部是否有癌”“区域淋巴结浸润情况”“是否浸润神经”“是否浸润血管淋巴管”“手术标本是否包含肌层”“变异组织学”“T分期”“N分期”
手术记录	“膀胱肿瘤是否单发”“膀胱肿瘤直径”“膀胱肿瘤位置”“TNM分期”“是否记录切除达肌层”
影像学记录	“是否为膀胱癌”“膀胱肿瘤直径”“肿瘤转移情况”“M分期”

表3 字段词典示例

Table 3. Example of field dictionary

字段名	字段值可能的描述
肿瘤浸润深度	**非浸润性**[非浸润型；非浸润性] **原位癌**[原位癌；原位尿路上皮癌] **浸润性**[浸润性尿路上皮癌；浸润型尿路上皮癌；浸润性尿路上皮乳头状癌；浸润性癌；浸润性乳头状尿路上皮癌] **上皮下结缔组织**[固有层内见癌；侵及固有层；固有层内可见癌；侵及黏膜固有层；固有层见极少许可疑浸润灶；侵及膀胱壁固有层；伴黏膜固有层局灶表浅浸润；固有层内见癌；侵及黏膜下层；侵及上皮下纤维结缔组织；固有层见癌浸润；固有层可见癌浸润；侵及膀胱黏膜固有层] **肌层**[固有肌层见癌；肌层组织内可见癌；侵及固有肌层；肌层可见癌；侵及膀胱壁部分肌层；侵及固有肌组织；可见固有肌层浸润；侵及膀胱肌层；侵及部分肌层；固有肌层组织内可见癌；侵及肌层；固有肌层内可见癌；浸润膀胱壁肌层；肌层组织内见癌] **浅肌层**[侵及浅肌层；侵及膀胱壁浅肌层；内1/2] **深肌层**[浸润至膀胱深肌层；侵及膀胱壁深肌层；侵及膀胱壁全层；侵及深肌浆膜；浸润深肌层；外1/2] **膀胱周围组织**[侵及膀胱壁浆膜纤维脂肪组织；侵及膀胱壁浆膜外纤维脂肪组织；侵及膀胱肌层外脂肪组织；侵及膀胱浆膜层；侵及膀胱周围纤维脂肪组织；侵及膀胱周围组织；侵及膀胱与前列腺交界处纤维脂肪组织]

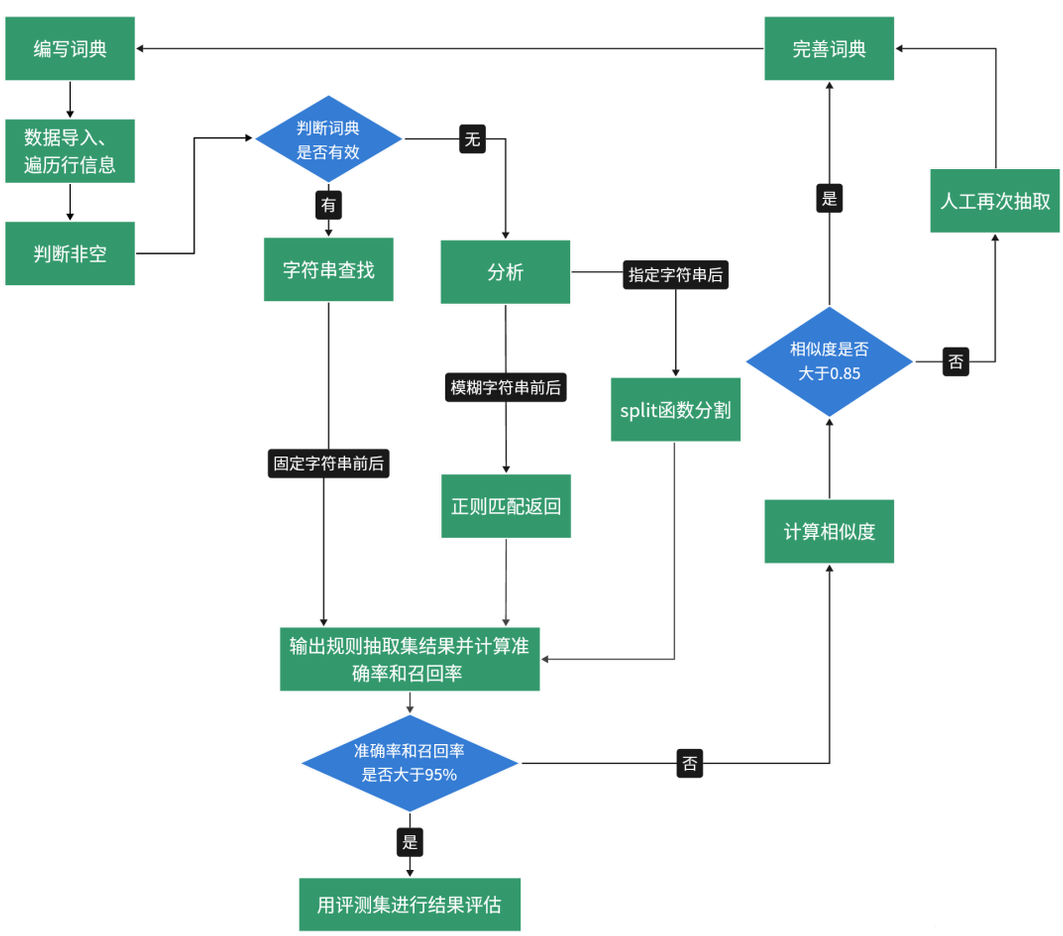

图1 抽取方法流程图

Figure 1. Flowchart of extraction method

名及操作。本文以“病理记录”工作簿（赋名为 df）的处理情况示例。读取“病理诊断”信息列（即目标字段所在列）信息后，对每一列分配出独立的代码块，利用 for i in range（len (df)）完成对 df 工作簿每一行的遍历工作。

步骤二，判断非空。某些信息列中存在信息缺失的情况，此时单元格的类型被判定为 float 类型，而 float 类型不支持有关字符串处理的内容。直接利用步骤一的方式遍历会报错，因此需要在进入遍历后且参与 if 语句判断之前，利用 if type（df [' 病理诊断 '] [i]）!=float 对信息列是否空白进行判断，非空白的行再进行下一步操作。

步骤三，判断词典是否有效并抽取结构化信息。包括以下几种情况：（1）若词典有效，可直接抽取出信息列中的字符串，如“分级”字段需要抽取出文本中的“高 / 中 / 低分化”“高 / 低级别”或“1/2/3 级”；或查找信息列中是否有特定的字符串，如“是否为膀胱标本”“是否包含肌层”等。（2）若词典无效：在本研究中，主要有以下两种情况：①需要“指定字符串后”处理的组合字段，如“区域淋巴结浸润情况”需要抽取出淋巴结的位置及对应的送检淋巴结与阳性淋巴结的数量，需要利用 split 函数识别关键词，由于该字段需抽取出不同位置淋巴结浸润数量，因此将抽取出来的数量存储到局部变量里，以便对不同区域的淋巴结进行描述。②需要“模糊字符串前后”处理的字段，通常无固定词组搭配，如“膀胱肿瘤直径”常在“新生物”周围出现，应抽取出“新生物”前后若干字符中含有的数字，即肿瘤的大小；或是在同一词组后跟随有干扰信息，如“入院记录”中的“饮酒时间”字段常跟在“饮酒史（年）”后，但该字符串后不仅有饮酒时间，还有判断是否饮酒的结果。利用 RE 中的 re 函数 re. findall（pattern, string, flags=0），可将所有匹配内容以列表形式返回。可以定义任意 pattern 为匹配模式，选择所要匹配的内容。

步骤四，输出规则抽取集结果计算准确率和召回率，计算相似度并完善词典。使用上述代码对规则抽取集进行字段抽取并输出结果，将该结果与人工抽取结果进行比对并计算其准确率和召回率。准确率 = 正确抽取的数量 / 已抽取到的数量，反映查准率；召回率 = 正确抽取的数量 / 待抽取的数量，反映查全率。

在根据规则抽取集抽取结果对词典进行完善以及后期根据新的语料补充词典时，需要人工对 RE 未能抽取出结果的文本进行抽取，该工作耗时耗力，故本研究利用自然语言处理方法，首先利用 jieba 库进行中文文本的分词工作 words= jieba. lcut（s），然后利用 Word2Vec 模型对每一个分词的结果转化为向量（Vector）。

for word in words:

v+=model [word]

接着将所有的 Vector 相加并求平均，得到整个句子的词向量（Sentence Vector）。

v/=len（words）

将关键词已经纳入词典的信息列与其余信息列比对，并对 Sentence Vector 夹角的余弦值计算相似度，挑选相似度高（＞ 0.85）且用 RE 无法抽取出结果的语句，用这些语句进行词典更新，可大大减少人工工作量。

审核规则抽取集准确率和召回率低于 95% 的字段，通过上述过程完善词典。同一个变量中不同的描述字段均有一定的相似性，因此通过计算词典中已有字段和词典中没有但信息列中存在字段的相似度，筛选出相似度较高的字段，由机器直接提取，录入词典；相似度不高的字段则通过人工再次提取的方式录取词典进行再提取。重复执行以上步骤直到目标字段的召回率和准确率高于 95%。

步骤五，采用评测集进行结果评估。使用经由人工抽取并加以完善后的词典对评测集字段进行抽取，计算其准确率和召回率。由于 EMR 数据存在稀疏性特点，部分字段只存在于个别患者的记录中，目标字段的缺失率将对抽取结果有显著影响。因此本研究设定了评测集目标字段缺失率的最大阈值为 40%，若人工抽取的结果显示该字段缺失率大于 40%，则将从数据源中重新抽取数据对原评测集进行补充，直至该字段缺失率降低至 40% 以下，以此作为新的评测集，进行结果评估。

1.3 代码示例及详解

以“区域淋巴结浸润情况（送检淋巴结位置、数量及阳性数量）”和“饮酒史：时间（年）”字段抽取为示例进行代码展示及解释，详见框 1。

（1）字段“区域淋巴结浸润情况（送检淋巴结位置、数量及阳性数量）”属于词典无效字段

```
#示例字段1
if '送检（左、右闭孔' in df['病理诊断'][i]:
    a = df['病理诊断'][i].split("（左、右闭孔")[1].split("、")[0]
    if '枚、' in df['病理诊断'][i]:
        b = df['病理诊断'][i].split("枚、")[1].split("枚")[0]
        if '淋巴结未见癌' in df['病理诊断'][i] or '淋巴结未见癌转移' in df['病理诊断'][i] or '淋巴结内均未见癌' in df['病理诊断'][i]:
            c = d = "0"
        if '淋巴结见癌' in df['病理诊断'][i]:
            c = a    d = b
        if "/" in a:
            c = a.split("/")[0]    a = a.split("/")[1]
        if "/" in b:
            d = b.split("/")[0]    b = b.split("/")[1]
df['区域淋巴结浸润情况（送检淋巴结位置、数量及阳性数量）'][i] = "左闭孔" + a + " " + c + "\n" + "右闭孔" + b + " " + d

#示例字段2
a = re.findall(r"饮酒史\(年\): (\d+)", str(df['个人史'][i]))
df['饮酒史：时间（年）'][i] = str(a)
```

框1 部分目标字段抽取代码示例

Box 1. Code examples for some target fields extraction

并且需抽取信息位于指定字符串后。这类字段有多种情况，此处选取一种送检（左、右闭孔）情况进行说明，使用 split 函数将“（左、右闭孔”后，“、”前的内容抽取出来，并存放在局部变量 a 中，一般为左闭孔淋巴结的数量，再通过分割将第一个“枚”字后和第二个“枚”字前的内容抽取出来，一般为右闭孔淋巴结的数量，存放在另一个局部变量 b 中，若有癌转移，且 a、b 变量中含有“/”时，前面的数字代表阳性淋巴结的数量，后面的数字代表送检淋巴结数量，用 split 函数将两个数字分别分割出来，存放进不同的局部变量。

（2）入院记录中的字段“饮酒史：时间（年）”属于词典无效字段并且需抽取信息位于模糊字符串后。为抽取“个人史”信息列中的饮酒年份，使用 re. findall 函数查找，并用 RE 规则 d^+，抽取出在“饮酒史（年）”后的所有数字以列表形式返回。

2 结果

在对评测集中缺失率过高的字段进行补充抽样后，本研究以“未删除缺失数据集”“删除缺失数据集”为评测集分别进行结果评估，若将缺失值认为是抽取结果之一，那么可得到包含所有正误对比的完整结果，但无法直观获得准确率和召回率对于字段词典的准确性、完整性以及 RE 抽取效果的不同体现，因为已抽取到的数量即为待抽取的数量，准确率和召回率的结果相同。如果删除缺失，即待抽取数量只包含有目标值的情况，区别于已抽取数量，将得到不一致的召回率和准确率从而对正则抽取效果进行评估，但是因为空值均被删除，无法反映实际为空却被 RE 抽取到了错误值的情况，因此同时报告两个数据集的结果，进行综合评估。结果详见表 4。

未删除缺失数据集的评估结果显示，病理记录中的大部分目标字段准确率和召回率均可达到 80% 以上，只有 3~4 个目标字段准确率和召回率低于 80% 但水平仍可达到 60%~80% 之间；手术记录中的“膀胱肿瘤是否单发”“膀胱肿瘤直径”“膀胱肿瘤位置”和影像学记录中“膀胱肿瘤直径”字段的准确率和召回率相对较低，在 60% 左右，手术记录和影像学记录中其它目标字段的准确率和召回率均可达到 95% 以上；入院记录的所有目标字段召回率与准确率均在 90% 左右。

删除缺失数据集的评估结果显示，大部分目标字段的准确率较未删除前有显著提升，可达到

表4 未删除缺失数据集与删除缺失数据集结果

Figure 4. Results of datasets without and with deleting missing data

抽取字段	未删除缺失数据集		删除缺失数据集	
	准确率	召回率	准确率	召回率
入院记录				
吸烟史：有/无	100.00	100.00	100.00	100.00
吸烟时间（年）	92.89	92.89	89.29	92.59
吸烟史: 平均（支/d）	100.00	100.00	100.00	99.25
戒烟：有/无	96.19	96.19	100.00	93.65
戒烟时间（年）	87.11	87.11	100.00	74.23
饮酒史：有/无	97.00	97.00	100.00	96.39
饮酒时间（年）	96.95	96.95	84.69	82.18
饮酒平均：平均（mL/d）	99.53	99.53	100.00	99.22
戒酒史：有/无	96.00	96.00	95.83	100.00
戒酒时间（年）	96.03	96.03	97.87	88.46
毒品接触史	100.00	100.00	100.00	100.00
病理记录				
是否为膀胱标本	99.00	99.00	99.00	99.00
是否为癌症	93.00	93.00	93.00	93.00
是否为膀胱尿路上皮癌	97.00	97.00	97.00	97.00
病理结果中是否记录明确分期	100.00	100.00	100.00	100.00
病理分级	99.00	99.00	98.39	100.00
肿瘤浸润深度	94.00	94.00	90.00	90.00
是否有输尿管残端、输精管断端癌浸润	63.50	63.50	95.92	39.17
是否有精囊腺、前列腺组织、子宫、阴道癌浸润	81.50	81.50	90.22	69.17
基底部是否有癌	55.40	55.40	100.00	25.78
区域淋巴结浸润情况	69.70	69.70	96.72	49.58
是否浸润神经	79.44	79.44	100.00	58.43
是否浸润血管淋巴管	74.49	74.49	100.00	52.38
手术标本是否包含肌层	77.13	77.13	98.59	61.95
变异组织学	71.19	71.19	100.00	37.80
T分期	93.00	93.00	89.83	88.33
N分期	88.00	88.00	88.00	88.00
手术记录				
膀胱肿瘤是否单发	50.34	50.34	81.25	14.94
膀胱肿瘤直径	61.36	61.36	58.11	40.95
膀胱肿瘤位置	69.66	69.66	81.67	56.32
TNM分期	99.30	99.30	100.00	97.78
是否记录切除达肌层	100.00	100.00	100.00	100.00
影像学记录				
是否为膀胱癌	94.00	94.00	94.00	94.00
膀胱肿瘤直径	60.59	60.59	52.33	36.89
肿瘤转移情况	96.00	96.00	96.00	96.00
M分期	96.00	96.00	96.00	96.00

95% 以上；但是召回率结果差异较大，病理记录中目标字段平均召回率在 75% 左右，手术记录中目标字段平均召回率约为 63%，入院记录中目标字段平均召回率为 94%，而影像学记录中除“膀胱肿瘤直径”召回率为 37% 外，其余字段召回率均高于 90%。

3 讨论

本研究结果显示，总体上基于 RE 方法抽取目标字段的准确率较高，说明人工总结的字段词典的查准率较高、准确性较强，原因可能是基于 RE 方法的信息抽取可精准匹配特定的文本模式。但是基于 RE 方法抽取目标字段的召回率相对较低且差异较大，反映出在部分字段上词典的查全率相对较低，完整性相对较差。原因可能有两种，一是由于规则抽取集不够具有代表性，导致人工总结的词典不够完整，后续可通过增加规则抽取集、迭代完善词典的方式解决；二是文本中目标字段对应的语言结构复杂、文本表述变化多样，RE 方法难以归纳概括全部规律而导致漏抽，查全率低，需要尽量全面总结字段表达规律以改善抽取结果。需注意的是，对高缺失率目标字段补充抽样时，受限于原数据，个别目标字段如“是否浸润神经”“戒烟时间（年）”等的补充抽样仍不能达到缺失率小于 40% 的目标，其评估结果可能随着包含目标值样本量的增大而波动。

既往有研究者以层叠条件随机场机器学习模型为基础，对包含入院记录、出院记录、辅助检查报告等非结构化文本的呼吸专科住院 EMR 进行了信息抽取，结果显示病历中各类文本信息抽取准确率和召回率分别为 92.12%、92.42% [11]。与本研究抽取结果对比发现，一方面，基于 RE 的信息抽取方法其准确率较高，对于大部分变量可以达到 98% 甚至是 100% 的准确率，但是召回率就显著逊色于基于机器学习的信息抽取方法，因为 RE 无法抽取规则集之外的文本信息；其次，对于表述简单的字段，如病理记录中的“是否为膀胱标本”“是否为膀胱尿路上皮癌”等字段，或半结构化文本中的字段抽取，如入院记录，RE 规则集的编写会相对简单，其抽取效果优于机器学习，可达到较高的准确率和召回率（98% 以上）。但是对于像影像学记录与手术记录中的“膀胱肿瘤直径”“膀胱肿瘤位置”等表述形式较为繁杂的非结构化文本，由于自由度过大、语法语义复杂，基于人工编写规则的信息抽取方法就难以较好归纳概括其表达规律，导致规则的编写过程耗时耗力，抽取的效果不佳。

本研究对于四类文本规则集归纳总结的时间耗时较久，侧面说明 RE 的人工依赖性较强。某些字段在文本中的出现频次低，其固有稀疏性导致研究者难以充分总结表达规律，而文本中有时出现的不规范数据输入情况也会直接影响规则集的总结和抽取结果的可靠性，并且在前期规则集的编写过程中需要大量的人工参与，而人工制定规则集的质量由文本的结构化程度决定并直接影响最终信息抽取效果。例如吴欢等的研究表明，针对语义简单、结构规范的文本，基于规则的模式匹配方法的信息抽取技术更简单、快速、易实现 [12]；对于像冠状动脉 CT 血管成像及钙化积分这类单病种且比较规范的报告，RE 是实现其结构化的最佳投入产出比方案，其制定的规则对于报告的结构和语言描述具有较高的依赖性 [13]。RE 始终是基于人工制定的规则进行信息抽取，但是由于 RWD 的多样性，可能会出现字符一致但是语境不同而导致抽取错误的情况，需要人工对表达式进行完善，也可能会影响该方法的适用、推广和维护。

需要注意的是，RE 规则集跨学科、跨单位、跨病种的可迁移性和可复用性较弱。在不同学科之间各类文本所涉及知识、术语、数据大相径庭以及 RE 对于规则集准确性有着高要求的前提下，对于某一学科中某一文本构建的规则集难以进行跨学科迁移。因不同机构、系统或医生记录习惯的不同，可能会导致无法实现规则集的大规模跨机构使用。对于围绕单一病种 EMR 记录构建的规则集，除入院记录中部分信息不具有疾病特异性，大部分记录如病理记录、手术记录、影像学记录等均具有疾病特异性表达，其规则集难以跨病种。

在对规则抽取集的词典进行革新迭代过程中，本研究采用了计算文本相似度的方法，通过计算无法抽取出信息或新纳入的文本与已归纳总结出规则集文本的相似度进行比对，从而提高词典更新的效率，一方面可减少人工工作量，提高信息抽取的召回率与准确率，另一方面则可更准确、高效地服务于新文本数据。该方法可以在一定程度上为上述 RE 规则集可迁移性、可复用性弱以及人工依赖性强的问题提供解决思路。

近年来，越来越多的研究者将目标聚焦于医疗文本中结构化信息的抽取，相关的方法与算法优化不断涌现。例如安辉对 RE 可视化编辑的实现，以降低 RE 的学习和使用难度[14]；相关研究者提出的基于文本表示的 RE 自动生成技术，可大大减少研究者概括规则集、撰写 RE 过程中耗费的时间以及人力资源成本[15]；同时 ChatGPT 类大型预训练语言模型的出现和发展，也为文本挖掘、信息抽取领域开辟了新的途径[16]。吴骋等积极探索新的多层次信息抽取模式，实现了对医疗文本中各种信息的多维解析与分类存储[17]。抽取方法的不断革新为医疗大数据的价值挖掘提供了有力抓手。

对于研究中发现的问题，可考虑以下解决方法：选择基于统计学方法的信息抽取方法，如机器学习、深度学习等，以大量样本数据为训练集进行模型训练从而实现对于非结构化数据的信息抽取[18]。已有研究者采用机器学习方法识别并抽取病历中药物滥用和药物使用障碍等相关信息[19]；从源头解决数据质量差、结构化程度低的问题，加强医院信息系统的顶层设计，树立医务人员对高质量数据价值的正确认识，规范医务人员对 EMR 等医疗数据的书写和核对，提高数据的结构化程度和质量。

本研究以 RE 为基础，针对膀胱癌 EMR 数据开展实践应用，具有一定的应用价值，但该方法存在一定局限性，诸如人工依赖性较强，部分字段抽取的准确率与召回率偏低等问题。并且本研究并未对 RE 规则集在跨病种、跨单位等的其他数据集上的抽取效果进行测试。后期研究团队将使用基于 Transformer 架构的深度学习模型对相同的数据进行信息抽取，并对比二者在操作流程、适用样本、构建时间、构建难度、抽取效率、抽取效果等方面的优劣，并纳入其他单位以及其他病种的 EMR 数据，以此为基础构建可视化平台，为研究者提供参考。

参考文献

1 US FDA. Real-world evidence program framework[EB/OL]. (2019-05) [2022-07-13]. https://www.fda.gov/drugs/webinar-framework-fdas-real-world-evidence-program-mar-15-2019.

2 杨羽，詹思延．上市后大数据药品安全主动监测模式研究的必要性和可行性[J]. 药物流行病学杂志，2016, 25(7): 401-404, 413. [Yang Y, Zhan SY. Analysis of necessity and feasibility in studies of post-marketing drug safety active surveillance based on big data[J]. Chinese Journal of Pharmacoepidemiology, 2016, 25(7): 401-404, 413.] DOI: 10.19960/j.cnki.issn1005-0698.2016.07.001.

3 阎思宇，李绪辉，陈沐坤，等．面向真实世界的知识挖掘与知识图谱补全研究（二）：非结构化电子病历信息抽取方法及进展[J]. 医学新知，2023, 33(5): 358-365. [Yan SY, Li XH, Chen MK, et al. Research on real-world knowledge mining and knowledge graph completion (II): methods and progress of information extraction from unstructured electronicmedical records[J]. Yixue Xinzhi Zazhi, 2023, 33(5): 358-365.] DOI: 10.12173/j.issn.1004-5511.202301016.

4 胡军伟，秦奕青，张伟．正则表达式在 Web 信息抽取中的应用[J]. 北京信息科技大学学报（自然科学版），2011, 26(6): 86-89. [Hu JW, Qin YQ, Zhang W. Regular expression and its applications to web information extraction[J]. Journal of Beijing Institute of Machinery, 2011, 26(6): 86-89.] DOI: 10.3969/j.issn.1674-6864.2011.06.019.

5 Cheung ATM, Kurland DB, Neifert S, et al. Developing an automated registry (Autoregistry) of spine surgery using natural language processing and health system scale databases[J]. Neurosurgery. 2023, 93(6): 1228-1234. DOI: 10.1227/neu.0000000000002568.

6 Flores CA, Figueroa RL, Pezoa JE. FREGEX: a feature extraction method for biomedical text classification using regular expressions[J]. Annu Int Conf IEEE Eng Med Biol Soc. 2019, 2019: 6085-6088. DOI: 10.1109/EMBC.2019.8857471.

7 范玉玲，顾进广，黄智生．中文医学指南的事件处理及其语义数据自动生成[J]. 中国数字医学，2015(9): 76-78, 112. [Fan YL, Gu JG, Huang ZS. Event handling of Chinese medical guide and the automatic generation of its semantic data[J]. China Digital Medicine, 2015(9): 76-78, 112.] DOI: 10.3969/j.issn.1673-7571.2015.09.026.

8 Humphrey PA, Moch H, Cubilla AL, et al. The 2016 WHO classification of tumours of the urinary system and male genital organs-part b: prostate and bladder tumours[J]. Eur Urol. 2016, 70(1): 106-119. DOI: 10.1016/j.eururo.2016.02.028.

9 Amin MB, Greene FL, Edge SB, et al. The eighth edition

AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging[J]. CA Cancer J Clin. 2017, 67(2): 93-99. DOI: 10.3322/caac.21388.

10 徐荣飞. Python 正则表达式研究 [J]. 电脑编程技巧与维护, 2015(9): 45, 49. [Xu RF. Python research on regular expressions[J]. Computer Programming Skills & Maintenance, 2015(9): 45, 49.] DOI: 10.3969/j.issn.1006-4052.2015.09.020.

11 梁立荣, 李长伟, 沈晔, 等. 基于层叠条件随机场模型的电子病历文本信息抽取 [J]. 计算机应用与软件, 2019, 36(10): 47-54, 112. [Liang LR, Li CW, Shen Y, et al. Text information extraction for electronic medical record based on cascaded conditional random field model[J]. Computer Applications and Software, 2019, 36(10): 47-54, 112.] DOI: 10.3969/j.issn.1000-386x.2019.10.009.

12 吴欢, 应俊, 王逸飞, 等. 乳腺癌病理文本的结构化信息提取 [J]. 解放军医学院学报, 2020, 41(7): 746-751. [Wu H, Ying J, Wang YF, et al. Structured information extraction from breast cancer pathological report texts[J]. Academic Journal of Chinese PLA Medical School, 2020, 41(7): 746-751.] DOI: 10.3969/j.issn.2095-5227.2020.07.022.

13 杨金荣, 喻杰, 叶豪, 等. 正则表达式在提取冠状动脉 CTA 和钙化积分报告结构化信息中的应用 [J]. 中国数字医学, 2022, 17(11): 38-44. [Yang JR, Yu J, Ye H, et al. Application of regular expression in extracting structured information of coronary artery CTA and calcification score reports[J]. China Digital Medicine, 2022, 17(11): 38-44.] DOI: 10.3969/j.issn.1673-7571.2022.11.008.

14 安辉. 健康评估中医学知识的可视化呈现与交互 [D]. 浙江: 杭州师范大学, 2019. [An H. Visual presentation and interaction of medical knowledge in health assessment[D]. Zhejiang: Hangzhou Normal University, 2019.]

15 王晓琳. 正则表达式生成与复杂正则表达式识别技术研究 [D]. 北京: 中国科学院大学, 2022. [Wang XL. Research on regular expression generation and complex regular expression recognition techniques[D]. Beijing: University of Chinese Academy of Sciences, 2022.]

16 鲍彤, 章成志. ChatGPT 中文信息抽取能力测评——以三种典型的抽取任务为例 [J/OL]. 数据分析与知识发现, 1-16. [Bao T, Zhang CZ. Extracting Chinese information with ChatGPT: an empirical study by three typical tasks[J/QL]. Data Analysis and Knowledge Discovery, 1-16. DOI: 10.11925/infotech.2096-3467. 2023.0473.

17 吴骋, 徐蕾, 秦婴逸, 等. 中文电子病历多层次信息抽取方法的探索 [J]. 中国数字医学, 2020, 15(6): 29-31. [Wu P, Xu L, Qin YY. Exploration on the multi-level information extraction method of Chinese electronic medical records[J]. China Digital Medicine, 2020, 15(6): 29-31.] DOI: 10.3969/j.issn.1673-7571.2020.06.009.

18 Adamson B, Waskom M, Blarre A, et al. Approach to machine learning for extraction of real-world data variables from electronic health records[J]. Front Pharmacol. 2023, 14: 1180962. DOI: 10.3389/fphar.2023.1180962.

19 周虎子威, 张云静, 于玥琳, 等. 机器学习方法在预测麻精药品不合理使用风险中的应用现状和思考 [J]. 药物流行病学杂志, 2023, 32(4): 446-457. [Zhou HZW, Zhang YJ, Yu YL, et al. Application of machine learning methods in predicting the risk of irrational use of narcotic and psychotropic drugs:current status and considerations[J]. Chinese Journal of Pharmacoepidemiology, 2023, 32(4): 446-457.] DOI: 10.19960/j.issn.1005-0698.202304010.

收稿日期：2023 年 08 月 03 日 修回日期：2024 年 02 月 22 日
本文编辑：桂裕亮 曹 越

引用本文：马文昊, 石涵予, 黄桥, 等. 面向真实世界的知识挖掘与知识图谱补全研究（三）：基于正则表达式对膀胱癌真实世界数据的结构化信息抽取[J]. 医学新知, 2024, 34(3): 312-321. DOI: 10.12173/j.issn.1004-5511.202308006
Ma WH, Shi HY, Huang Q, et al. Research on real-world knowledge mining and knowledge graph completion (III): structured information extraction from real world data of bladder cancer based on regular expression[J]. Yixue Xinzhi Zazhi, 2024, 34(3): 312-321. DOI:10.12173/j.issn.1004-5511.202308006