

面向脑卒中防治知识图谱的风险评估与分类^{*}

严 越

郑 静 林德南

吕 果 倪士光 袁克虹

(清华大学深圳国际研究生院
深圳 518055)

(深圳市医学信息中心
深圳 518000)

(清华大学深圳国际研究生院
深圳 518055)

〔摘要〕 分析脑卒中筛查评估现状,构建以危险因素为核心的可视化脑卒中防治知识图谱,阐述其概念结构及构建过程、风险评估模型架构及因子系数计算方法、结果分析与危险因素分类等。

〔关键词〕 脑卒中;知识图谱;文本挖掘;危险因素;风险评估

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.09.006

Risk Assessment and Classification for Knowledge Graph of Stroke Prevention and Treatment YAN Yue, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China; ZHENG Jing, LIN Denan, Shenzhen Medical Information Center, Shenzhen 518000, China; LV Guo, NI Shiguang, YUAN Kehong, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China

〔Abstract〕 The paper analyzes the status quo of stroke screening and evaluation, builds a visual knowledge graph of stroke prevention and treatment with risk factors as the core, and elaborates its conceptual structure and building process, risk assessment model structure and factor coefficient calculation method, result analysis and risk factor classification, etc.

〔Keywords〕 stroke; knowledge graph; text mining; risk factor; risk assessment

1 引言

脑卒中是常见脑血管疾病,俗称“中风”,由脑部血管突然破裂或血管阻塞引起,对应国际疾病分类 ICD-10 标准编码号为 I64.X04。从全球范围看,我国卒中终生风险及卒中引发疾病负担位居世界第 1,风险高达 39.3%^[1]。从国内角度看,脑卒

中是我国成年人致死、致残首位病因,具有发病率高、致残率高、死亡率高、复发率高、经济负担高的特点,伴发病年轻化、普遍化等不良趋势^[2]。病患确诊脑卒中后,通常只能控制难以根治。因此需要充分发挥基层医护人员作用,做好评估筛查,有效规避卒中风险,做到“早预防,早治疗”^[3]。但是脑卒中危险因素预测与及时监控困难,基层医护人员对此缺乏工具。针对脑卒中高危人群的筛查常用 ABCD2,美国国立卫生研究院卒中量表(National Institute of Health Stroke Scale, NIHSS),Rankin, Barthel 量表等^[4],但复杂、耗时耗力;FAST 等评价方法^[5]较简单但难以满足需求。目前尚缺乏对我国或亚洲人种族危险因素综合排序研究,也未定量考虑危险因素间的耦合关系。

〔收稿日期〕 2020-02-19

〔作者简介〕 严越,硕士研究生,发表论文 5 篇;通讯作者:郑静,博士,高级工程师。

〔基金项目〕 深圳市三名工程项目(项目编号:No.SZSM201811094)。

2 脑卒中防治知识图谱可视化构建

2.1 构建方法

2.1.1 总体思路 首先对相关实体和关系进行定义，综合数据源给出第1层（顶层）实体节点；其次在领域知识帮助下归纳抽取顶层实体之间关系；再次对每个实体向下抽取节点，用共词分析法^[6]、半监督机器学习方法^[7]进行文本挖掘以确定节点，直到能够获取实体属性为止^[8]。脑卒中防治医学节点（Prevention Entity, PE）指脑卒中疾病在防治方面可唯一标识的医学实体或属性，通常出现在文本数据、电子病历、调研结果表单等各种渠道中。实体节点的父节点必须为实体或主题节点，子节点可为实体或属性。脑卒中防治医学关系（Prevention Relationship, PR）表示不同脑卒中防治医学实体或属性间发生的事实关系，可表示为 PR（PE_i, PE_j）或 PR（PE_i, PA_i），其中 PE_i, PE_j为脑卒中防治医学实体，PA_i为脑卒中防治医学属性。

2.1.2 概念层构建 首先定义顶层主题词、整合

并筛选高频主题，总结出6方面主题词并将其作为脑卒中防治知识图谱概念层的顶层实体节点，分别为患者实体、危险因素实体、筛查手段实体、健康指导实体、症状表现实体与预防方式实体。将危险因素作为知识图谱核心实体，整理出9种脑卒中防治医学关系类型：has、instance of、attribute of、take、control、influence、improve、test、lead。大致可分为跨层与同层两种，见表1、图1。

表 1 9 种脑卒中防治医学关系

关系类型	关系表示	关系含义
跨层	A instance of B	实体 B 是实体 A 的实例或对象
跨层	A attribute of B	实体 A 是实体 B 的属性值
同层	has	患者实体拥有危险因素实体
同层	test	筛查手段实体能检测并筛选危险因素实体
同层	influence	健康指导实体会影响危险因素实体
同层	lead	危险因素实体会导致早期的症状表现实体
同层	take	患者实体可以采用健康指导实体
同层	improve	健康指导实体将会改进症状表现实体
同层	control	预防方式实体一定程度上可控制危险因素实体

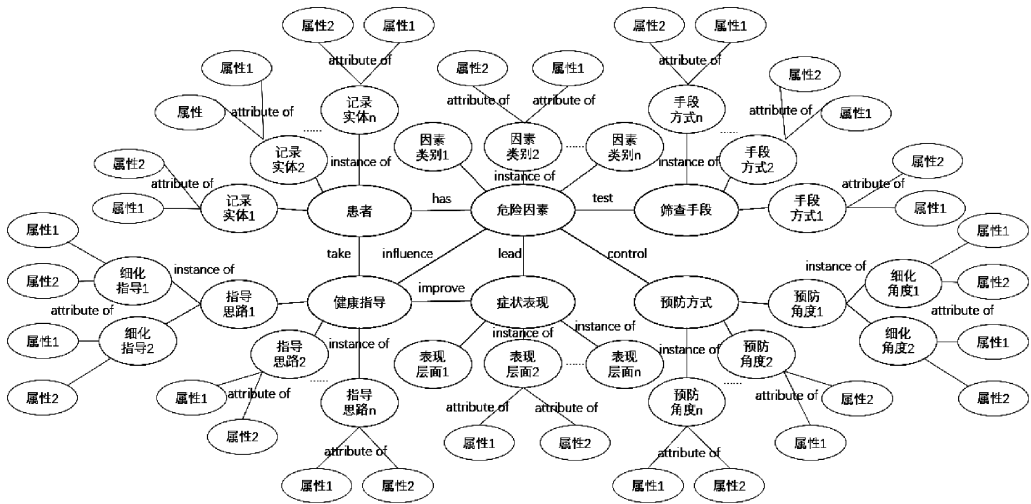


图 1 脑卒中防治知识图谱概念层结构

2.1.3 节点提取路径 除顶层外的节点都需由文本挖掘识别与命名。针对不同数据类型采取不同处理方式，见图2。对文献采用共词分析法^[6]挖掘危险因素实体及内在联系；对病例和调研结果等采用

表单解析法^[9]挖掘实体。最后对挖掘到的实体进行汇总与筛选，逆向剔除不合语言规则的危险因素^[10]并请专家把关。

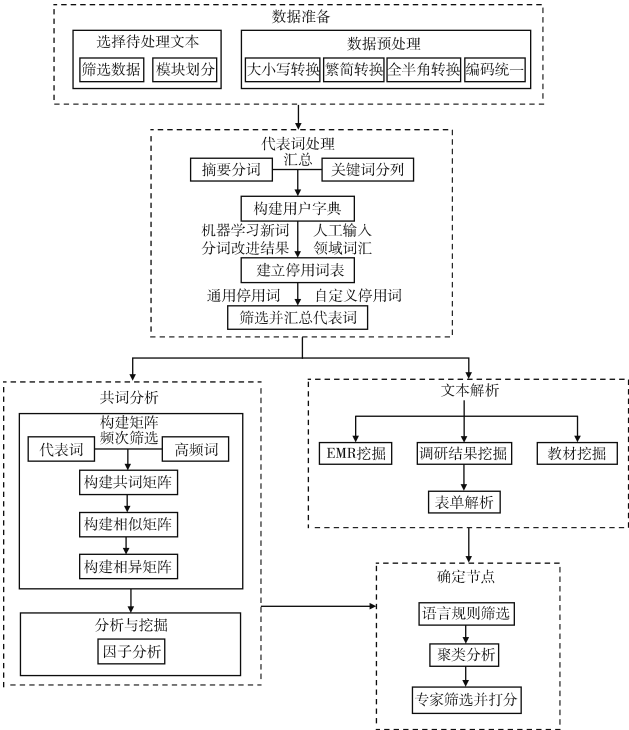


图2 节点提取实现路径

2.2 构建过程

2.2.1 数据预处理 (1) 数据采集。为保证图谱内容专业全面，数据来源涵盖电子病历与调研结果、学术论文与医疗教材、政府报告与规范指南3类。采用爱爱医专业医学平台获取有效病例54则，中国知网（China National Knowledge Infrastructure, CNKI）获取学术期刊文献1222篇，调研有效问卷数量共计257份，获取人民卫生出版社《神经病学》第8章脑血管疾病以及医生培训的脑卒中症状识别及院前急救教材6份，近3年中国脑卒中防治报告及14份规范指南。(2) 数据清洗。用文本整理器整合格式、实现大小写与全半角转换等。构建用户字典使摘要分词结果更准确^[11]，共导入机器和手工汇聚形成的406条用户字典帮助机器进行分词。爬取1208条常用中文停用词，通过机器学习最终建立包含1555条词条的停用词表库，剔除对挖掘无意义的词语^[12]。

2.2.2 挖掘特征词 (1) 构建共词矩阵。清洗后得到474个待处理词，词频筛选后得到249条高频词。采用Ochia系数将每个数值都除以与之相关

的两个关键词总频次开放的乘积，用Transport函数进行转秩，得到相似矩阵。为降低统计误差，将危险因素相似矩阵转化为表示两词间差异程度的相异矩阵，数值大说明代表词间距离大、相似度低。

(2) 文本挖掘。采用自下而上的合成聚类初步判定代表词间关联程度，再用主成分分析法确定最少能反映原始数据的公共因子，碎石图结果保留45个特征值大于1的因子。对旋转后的主成分因子载荷系数矩阵调用Abs函数求其绝对值，筛选大于0.5对命名有帮助的主成分命名因子。

(3) 筛选及输出候选特征词。因疾病及药物医学术语词性构成规则难以总结，故采用逆向语言规则法^[10]，过滤不能作为危险因素实体的噪音词语，如方位词、连接词等。以可干预危险因素为例的部分节点梳理结果，见表2。

表2 以可干预危险因素为例的部分节点梳理

顶层节点	第2层	第3层	更细层次节点
危险因素	可干预因素	疾病因素	高血压、糖尿病、血脂异常、心脏病、神经病变、感染等
		生活方式	吸烟、酗酒、缺乏运动锻炼、滥用药物、生活不规律等
		其他因素	居住地、手术、激素、心理状态、近亲结婚、胎盘早剥等

2.2.3 存储与可视化 Neo4j在乳腺癌等医学知识图谱^[13]的绘制中已被验证适用于生物医学领域，数据存储到Neo4j后共得到节点165个、关系168条。关键危险因素权重以属性方式存储，见图3。

3 基于脑卒中防治知识图谱的风险评估模型

3.1 模型架构

3.1.1 多可干预危险因素耦合作用模型 脑卒中疾病规律尚未完全探明，依赖作用的未知性增大致病因素模糊度。利用图谱所挖掘的典型危险因素，构建多可干预危险因素耦合作用模型，见图4。模型从疾病因素、生活方式和其他方面因素考虑子系统之间相互作用。

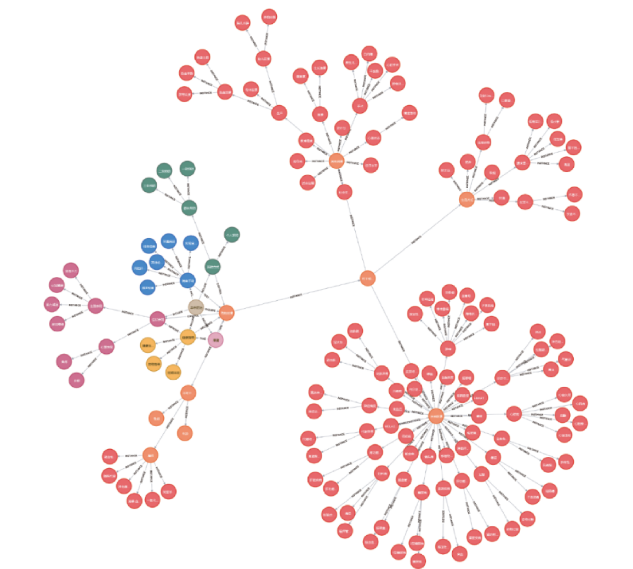


图3 知识图谱整体可视化效果

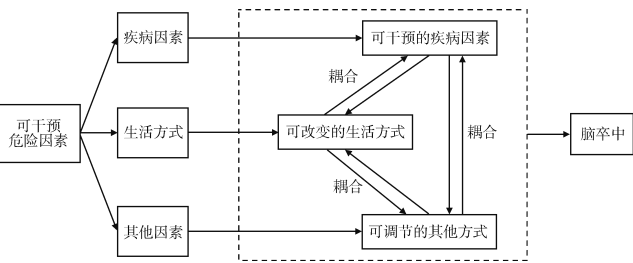


图4 多可干预危险因素耦合作用模型

3.1.1.2 影响因素 从知识图谱节点中归纳出表现力好、影响突出的9个因素：在生活方式层次下的节点有F₁₁（吸烟）、F₁₂（酗酒）、F₁₃（饮食不当）；疾病因素层次下的节点有F₂₁（高血压）、F₂₂（糖尿病）、F₂₃（心脏病）；其他因素层次下的节点有F₃₁（心理状况）、F₃₂（经济状况）、F₃₃（教育程度）。

3.1.1.3 相互作用矩阵 在矩阵主对角线上放置所有脑卒中风险影响因子，其值表示该风险因子会对脑卒中风险产生影响的程度。多个影响因子相互耦合作用则放置在次对角线位置，其值表示因素之间耦合作用对我国居民患脑卒中风险的影响程度，见图5。

$$P=\begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots & P_{1,m} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots & P_{2,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots & P_{i,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{m,1} & P_{m,1} & \cdots & P_{m,j} & \cdots & P_{m,m} \end{pmatrix}$$

图5 相互作用矩阵

3.2 因子系数计算方法

关系矩阵中的 $P_{i,j}$ 表示第*i*个风险影响因子单独对脑卒中致病的影响，根据式（1）-（3）计算影响整体或个体脑卒中风险性的第*i*个影响因子权重*k*。其中，*n*表示脑卒中风险的主要影响因子数量； $S_R(i)$ 表示风险因子*i*自身风险系数和与对其他影响因子产生耦合作用的系数之和； $S_C(i)$ 则表示风险因子*i*自身风险系数和与其他因子对*i*因子产生耦合作用的系数之和。再据式（4）进一步计算整体或个人的可控危险因素程度。 V_i 表示某个脑卒中风险因子*i*参数的值： $U_i = P_i, i / (\max - \min)$ 。 SD 是根据多个影响因子计算得到的用于评价脑卒中患病风险性的无量纲数值，称为可干预因素下的脑卒中评价系数^[14]。

$$k_i = \frac{S_R(i) + S_C(i)}{2 \sum_{i=1}^n P_{i,j}} \quad (i=1, 2, 3, \cdots, n) \quad (1)$$

其中：

$$S_R(i) = \sum_{i=1}^n P_{i,j} \quad (i=1, 2, 3, \cdots, n) \quad (2)$$

$$S_C(i) = \sum_{i=1}^n P_{j,i} \quad (i=1, 2, 3, \cdots, n) \quad (3)$$

最后：

$$SD = \sum_{i=1}^n K_i V_i \quad (0 \leq SD \leq 1) \quad (4)$$

3.3 求解因子系数示例

采用专家半定量取值法分别对节点间不对等耦合作用打分，获得9×9相互作用矩阵。按照可控危险因素影响作用大小和强烈程度划分为5级，采用无量纲数值0-4来定量表达无影响、弱影响、中等影响、强影响和极强影响作用。通过医院专家打分，每个数值分别表示可控危险因素*i*作用于可控危险因素*j*而对整体脑卒中患病风险产生的影响。计算影响整体脑卒中患病风险性的第*i*个可控危险因素的权重*k*，见表3。根据式（4）得到居民总体SD值为0.044 22，属中等风险范围，可作个人脑卒中风险的基准参考。构建模糊层次分析模型不仅能对群体情况作出判断，还能预测个人脑卒中风

险，将群体或个人脑卒中评价系数 SD 值等距分为 5 个等级，依照等级规范及时采取防治措施。

极弱风险、弱风险、中等风险、强风险和极强风险

表 3 脑卒中可干预危险因素权重及排序

排序	F_{ij}	危险因素名	$S_R(i)$	$S_C(i)$	$S_R(i) + S_C(i)$	权重 k
1	F_{31}	心理状况	26	27	53	0.066 25
2	F_{21}	高血压	18	30	48	0.060 00
2	F_{23}	心脏病	19	29	48	0.060 00
2	F_{12}	酗酒	33	15	48	0.060 00
5	F_{22}	糖尿病	18	29	47	0.058 75
6	F_{11}	吸烟	27	15	42	0.052 50
7	F_{13}	饮食习惯	23	16	39	0.048 75
7	F_{33}	教育程度	20	19	39	0.048 75
9	F_{32}	经济状况	18	20	38	0.047 50

4 结果分析与危险因素分类

4.1 极值分析

4.1.1 最易影响其他可控危险因素 即酗酒，不仅其自身对于脑卒中有极强影响作用，还会诱发高血压、心脏病等严重疾病，而这些疾病本身也是脑卒中发病的重要危险因素，故会造成负面效果极强的多因素耦合作用。

4.1.2 最易被其他因素所影响因素 即高血压，故高血压防控需要从多方面入手。心理状况是作用最明显的脑卒中可控危险因素，一方面其对吸烟、饮酒等不良生活习惯因素产生作用，另一方面脑卒中急性发病与此高度相关。从数据与病例中屡次出现的“情绪激动时起病”可知不能忽视心理健康，建议学会调节情绪、降低压力，在需要时寻求情感支持。

4.2 危险因素分类

从相互作用角度将危险因素分成 3 类： $SR > SC$ 、 $SR \approx SC$ 、 $SR < SC$ 。（1） $SR > SC$ 类因素。对其他脑卒中可控危险因素的影响明显大于其被别的因素所影响，是脑卒中防治的真正源头。主要包括酗酒（33:15）、吸烟（27:15）、饮食习惯（23:16），多为生活习惯，故对这类危险因素做好防控有重大意义。（2） $SR < SC$ 类因素。易被其他卒中可控危险因素影响，主要包括高血压（18:30）、心脏病（19:29）、

糖尿病（18:29），多为疾病。这类因素往往是其他因素所导致的结果，应引起重视，采取实际手段规避脑卒中风险，如服用药物等。（3） $SR \approx SC$ 类因素。影响其他危险因素和其自身被影响的作用相当，主要包括心理状况（26:27）、教育程度（20:19）、经济状况（18:20）。这类因素是危险因素的中间环节，既受其他因素影响同时也会影响其他因素。从长远看改善这类危险因素可阻断疾病发生链条。

5 结论

5.1 主要研究成果

本文初步构造面向脑卒中防治的知识图谱，集成多种疾病相关知识资源，有效地将庞杂数据资源转化为结构化、可视化知识。进而提出一种便于推广的风险预测模型，综合模糊层次分析法和耦合理论优点，能综合考虑、定量描述危险因素相互作用关系，提出将危险因素按 SR 与 SC 分类，而非单方面强调对疾病因素管控。面向知识图谱构建的风险评估模型可成为普通群众自查风险的有效工具和基层医护人员诊断的辅助工具。

5.2 未来研究方向

知识图谱逐步与医疗领域结合，能够解决当前医疗电子数据庞杂、知识管理困难的问题。国内乳腺癌、乙肝等少数疾病已经开始尝试构建知识图谱^[13,15-17]，但构建方法与标准不统一、真实数据应

用较少^[18-19]。中文医学知识图谱起步较晚,在覆盖疾病种类、数据质量等方面仍有发展空间,应加强与国外医学图谱的连接^[20]。脑卒中等高危疾病研究有待加强^[21],需进一步探索医学知识的自动提取技术^[9]。应发挥知识图谱精准快速查询普及防治知识的优势,成为脑卒中互答、诊断辅助等系统的基础^[22-23],服务于基层临床评估。脑卒中发病年轻化、普遍化等不良趋势逐渐显现,快速节奏与压力对青壮年人群生活方式产生的负面影响需加强重视。各年龄阶段人群均需做好脑卒中防治工作,在群体基础上针对个人情况进行调整,从根源进行防治。

参考文献

- GBD 2016 Causes of Death Collaborators. Global, Regional, and National Age - Sex Specific Mortality for 264 Causes of Death, 1980-2016: a systematic analysis for the global burden of disease study 2016 [J]. Lancet, 2017, 390 (10100): 1151 - 1210.
- 中华人民共和国国家卫生健康委. 2018 中国脑卒中防治工程工作报告 [EB/OL]. [2020-01-05]. http://slfsw.h.xiangzhan.com/p - article _ detail/id - 128150.html.
- 国家卫生健康委员会. 2018 中国卫生健康统计提要 [M]. 北京: 中国协和医科大学出版社, 2018.
- 段洪连, 刘美云, 张拥波, 等. 缺血性脑卒中常用评估量表及其最新研究进展 [J]. 中国全科医学, 2011, 14 (35): 4018 - 4021.
- 柯颖华. 门诊使用 FAST 评分量表对可疑脑卒中患者快速评估的价值研究 [J]. 世界最新医学信息文摘, 2019, 19 (82): 20 - 21.
- Zhenyong Wu, Jihua Liao, Wenyan Song, et al. Semantic Hyper - graph - based Knowledge Representation Architecture for Complex Product Development [J]. Computers in Industry, 2018 (100): 43 - 56.
- 江兵兵. 基于贝叶斯方法的半监督学习算法研究 [D]. 合肥: 中国科学技术大学, 2019.
- 崔雷. 生物医学文本挖掘: 步骤与工具 [J]. 中华医学图书情报杂志, 2017, 26 (3): 1 - 5.
- Ford E, Carroll J A, Smith H E, et al. Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: a systematic review [J]. Journal of the American Medical Informatics Association, 2016, 23 (5): 1007 - 1015.
- 胡婧, 刘伟, 马凯. 基于机器学习的高血压病历文本分类 [J]. 科学技术与工程, 2019, 19 (33): 296 - 301.
- 温雯, 伍思杰, 蔡瑞初, 等. 面向专业文献知识实体类型的抽取和标注 [J]. 中文信息学报, 2018, 32 (1): 102 - 115.
- Alex Gerage. New Framework Brings Accuracy, Efficiency to Identifying Stop Words [EB/OL]. [2019 - 12 - 02]. <https://techxplore.com/news/2019-12-framework-accuracy-efficiency-words.html>.
- 崔洁, 陈德华, 乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究 [J]. 计算机应用与软件, 2017, 34 (12): 122 - 126.
- 乔万冠, 李新春. 多因素耦合作用下煤矿企业风险评价 [J]. 煤炭工程, 2014, 46 (4): 145 - 148.
- 刘俊丽, 张秀梅, 蒋勇青. 基于文本挖掘的乙型肝炎相关文献知识图谱分析 [J]. 医学信息学杂志, 2014, 35 (1): 48 - 53.
- 喻文茜, 陶红兵, 金莉. 基于知识图谱的肿瘤 MDT 研究可视化分析 [J]. 中国医院管理, 2017, 37 (9): 45 - 47.
- 马晨浩. 基于甲状腺知识图谱的自动问答系统的设计与实现 [J]. 智能计算机与应用, 2018, 8 (3): 102 - 107.
- Rajkomar A, Oren E, Chen K, et al. Scalable and Accurate Deep Learning with Electronic Health Records [J]. NPJ Digital Medicine, 2018, 1 (1): 18.
- Jinghe Zhang, Jiaqi Gong, Laura Barnes. HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records [C]. Philadelphia: 2017 IEEE/ACM International Conference on Connected Health: applications, systems and engineering technologies (CHASE), 2017.
- Paulheim H. Data - Driven Joint Debugging of the Bpedia Mappings and Ontology [C]. Cham: European Semantic Web Conference, 2017: 404 - 418.
- Jeyaraj D Pandian, Seana L Gall, Mahesh P Kate, et al. Prevention of Stroke: a global perspective [J]. Lancet, 2018, 392 (10154): 1269 - 1278.
- Wu J, Liu X, Zhang X, et al. Master Clinical Medical Knowledge at Certificated - doctor - level with Deep Learning Model [J]. Nature Communications, 2018, 9 (1): 4352.
- 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用 [J]. 计算机研究与发展, 2018, 55 (12): 2587 - 2599.