• 医学信息研究 •

医学知识库语言学特征比较分析*

孙月萍 侯 震 侯 丽 李 姣

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 以美国国家癌症研究所面向医师和患者提供的癌症综合信息库（Physician Data Query，PDQ®）为例，分别从词汇和句法角度对 PDQ 的专业版和公众版进行对比分析，了解面向患者与面向医师的知识库的显著统计差异。

〔关键词〕 医学知识库；癌症综合信息库；语言学特征；术语；句法分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.01.011

Comparative Analysis of Linguistic Features in Medical Knowledge Base *SUN Yue－ping, HOU Zhen, HOU Li, LI Jiao, Institute of Medical Information Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Taking Physician Data Query (PDQ®), which is provided for doctors and patients by National Cancer Institute (NCI), for example, the paper carries out comparative analysis on professional and public editions of PDQ from the perspective of vocabulary and syntax to make clear the significant statistical difference between the knowledge bases facing patients and doctors.

〔**Keywords**〕 Medical knowledge base; Physician Data Query (PDQ); Linguistic features; Term; Syntactic analysis

1 引言

在医学领域不同受众尤其是医师和患者，有不同的知识需求。有研究表明为使患者的知识、动机和实践符合医学指南，需要加强糖尿病患者关于医学指南（Guidelines）的教育工作[1]。医师和患者对国家级医学综合信息所持有的知识以及需求都有很大区别[2]。在国外医学知识库建设中，已有很多工作同时为医师和患者提供服务。英国医疗委员会（General Medical Council）于 2010 年、2013 年发布的优秀医疗实践（Good Medical Practice），不仅为医师提供医疗实践指南，同时也为患者发布了关于患者对医师的期望为主题的指南[3]。自 2002 年 10 月起美国国家癌症研究所发布了癌症综合信息库（Physician Data Query，PDQ®），分别为医师和患者提供了不同版本[4]，该资源被广泛认可为针对不同

〔修回日期〕 2017－09－30

〔作者简介〕 孙月萍，助理研究员，博士；通讯作者：李姣，副研究员。

〔基金项目〕 国家社会科学基金“面向知识服务的公众健康知识组织体系构建研究”（项目编号：14BTQ032）；中国医学科学院中央级公益性科研院所基本科研业务费项目“中文生物医学开放式概念关系抽取研究”（项目编号：2016RC330005）；中央级公益性科研院所基本科研业务费项目“生物医学领域大规模语义计算关键技术研究”（项目编号：2016ZX330010）。

受众的重要癌症信息和教育资源[5]。国内医学指南的知识库构建工作相对落后。李楠等[6]研究表明我国指南制定水平与国际水平有较大差距，主要体现在制订的严谨性和编辑的独立性。为保障以医学指南为代表的医学知识的严谨性，除重视证据以及证据质量的分级，还应重视对患者偏好和价值观的收集[7]，即重视医师和患者的知识背景、应用背景的差异。

为探讨成熟的医学指南知识库如何实现从集中最佳医学证据的指南到为患者提供科普教育资料的转化，本研究尝试以癌症综合信息库 PDQ 为例，从计算语言学角度来解析该问题。结合应用背景，选择以下角度解析该转化问题：（1）术语统计角度。在实际应用中，面向医师和患者的资源建设首先要解决的问题是术语问题。一方面患者往往很难理解专业医师提供的用药或者疾病治疗信息，另一方面医师有时也会在解读患者的描述时产生疑惑。医学术语的来源比较广泛，包括医学研究和实践、健康科学、药学和治疗产品等，因此医学术语号称科学界的“语言学丛林”，一个多来源混合术语栖息地。对于患者和看护，由于其医学背景有限，对医学术语的理解往往与专业医师存在差距。考虑到术语对于可读性和可理解性的重要性，基于术语统计工作研究专业版和公众版的显著差异是一个可行的切入点。（2）构句统计角度。为进一步了解不同版本医学指南的语言风格差异，统计句长分布并进一步统计构句成分分布。针对某一种癌症信息的不同版本，除术语不同，在具体内容的诠释方式上是否也存在差异，本研究将该差异具体落实到构句成分（词性标注）的统计差异。其中，构句成分指句子分词结果中的每个单词的词性，包括名词、动词、形容词或其他词性。Peters 等[8]对在线癌症信息的语言、术语和可读性进行了研究，使用基于词汇统计的方法对比分别针对医师和公众的乳腺癌信息的词汇信息，发现针对不同的阅读等级，词频排序和关键性值具有显著差异。但该工作以字（Word）为分析单元，没有考虑到不同对象的术语（Term）使用差异。前期已有研究分别对比了面向医师的专业版 PDQ 和面向患者的公众版 PDQ 的结构性特征[9]，尚未有研究利用术语和词性标注统计结果诠释 PDQ 癌症信息不同版本的差异。综上所述，研究选择语言学特征角度，以癌症综合信息库 PDQ 为研究对象，解析医学指南知识库从集中了最佳医学证据的指南到为患者提供科普教育资料的转化问题。参考已有研究，本研究目标是从构词和构句角度来量化比较面向医师和患者的 PDQ 癌症信息，以期为医学教育资源转化编辑工作提供量化评测指标。

2 研究方法和研究内容

2.1 数据来源

本研究选取 PDQ 发布的 2014 年第 28 周癌症综合信息，从中抽取了 320 英文语料，按照其版本（受众）拆分为两个数据集，作为平行语料。其中的 160 篇受众为医师，另外的 160 篇受众为患者。

2.2 研究方法

（1）构建一个平行 PDQ 数据集，包括对齐的癌症综合信息（专业版与公众版）。（2）基于开放获取协作（Open Access Collaboration，OAC）消费者健康词典（Consumer Health Vocabulary，CHV）匹配的词汇进行统计分析，包括文本术语、CHV 术语以及医学一体化语言系统（UMLS）术语的覆盖统计和难度分数等。（3）基于句法的统计分析，包括句子数、句子长度以及词性等。（4）基于词汇和句法的统计分析结果，为量化医学教育信息编辑提供指南。具体分析流程，见图 1。

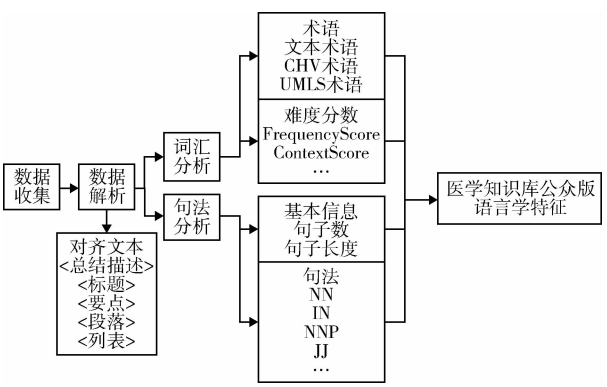

图 1 医学知识库语言学特征比较分析流程

2.3 研究内容

基于词汇的统计分析主要通过术语词典完成。OAC CHV[10]从查询日志中抽取健康相关词或词组，例如心脏病等，将这些词或词组关联到健康护理专家使用的专业词汇。利用该词典可以将"exanthema"（疹病）这样的专业词汇“转化”为"rash"（疹病）。在 OAC CHV 中，每个词汇都具有唯一的标识 CHV_ string_ id，如果为概念词汇，则用 CHV_ concept_ id 标识，第 1 列为 UMLS 概念 ID（CUI），第 2 列为术语，第 3 列为术语释义，另有一些标记，包括词汇是否是用户倾向使用词汇，是否为 UMLS 倾向使用词汇以及词汇是否可忽略，可忽略的词或者是拼写错误，或者有其他异常。此外每个词汇有一个特定的词汇难度分数，表示词汇理解的难易程度。CUI 分数来源于当前概念与已知易懂和难懂词汇的相关度，Combo 分数的取值来自于词汇难度分数与 CUI 分数的结合。本研究选取最新的 OAC CHV 词典，分别查看从文本中抽取的术语（Term）、CHV 名称（CHV Preferred Name）和 UMLS 名称（UMLS Preferred Name）在两个版本的覆盖率并计算术语的难度分数，作为重要的语言学特征。另外为了更深程度地考察专业版和公众版的差异，基于语言学方法对两个版本的癌症综合信息进行了平均句子个数和平均句子长度分析，基于句法分析工具[11]，对每个句子进行句法分析，考察各词性的统计差异。

3 研究结果

3.1 词汇

基于 OAC CHV 词典，本研究设计了基于词汇的统计分析实验，比较专业版和公众版的词汇覆盖率（文本术语、CHV 术语、UMLS 术语）以及词汇的难度分数总计，见表 1。由于专业版的篇幅普遍大于公众版，如果单纯按照词频总计比较，不能排除因篇幅差异带来的不一致性。因此在实验中，当 1 个词汇在 1 篇癌症信息总结中出现 1 次或多次时，其频数统计计为 1 次。此外，考虑到本研究的重点是考察两个版本构词的差异性，在使用 OAC CHV 词典进行统计之前，去除了词典中 CHV 术语（CHV Preferred）和 UMLS 术语（UMLS Preferred）词形一致的条目。统计结果表明，去除篇幅的影响后，专业版词汇中术语量均明显高于公众版。假设以文本术语（TermInText）在两个版本的覆盖率比例 1.65 为基准，公众版的 CHV 术语词汇明显高于专业版，而其 UMLS 术语词汇明显低于专业版，说明公众版在用词时更倾向于公众倾向使用的 CHV 术语。难度分数结果表明专业版的难度分数均高于公众版。难度分数比率越高，代表公众版的难度分数越低。基于词频的难度分数 FrequencyScore、已知易懂和难懂词汇的相关度的 CUIScore 以及两个 ComboScore 的对比值无明显差异。基于上下文计算的 ContextScore 对比尤为显著。说明在版本转化过程中，编辑人员充分考虑了词的理解难度问题。

表 1 基于 OAC CHV 的不同版本癌症综合信息术语覆盖对比

统计项	专业版	公众版	比率（专业版/公众版）
文本术语	113 480	68 794	1.65
CHV 术语	218 037	149 333	1.46
UMLS 术语	142 551	78 459	1.82
难度分数（ContextScore）	85 396.16	54 058.09	1.58
难度分数（FrequencyScore）	126 224.38	91 836.14	1.37
难度分数（CUIScore）	209 024.93	145 856.26	1.43
难度分数（ComboScore_ Pro）	231 370.23	162 267.62	1.43
难度分数（ComboScore_ nsw_ Pro）	220 240.51	151 801.08	1.45

3.2 句法

去除元素标记后对文本进行分句，分别统计公众版和专业版的句数以及长度并进一步对各个句子进行句法分析，取得结果，见表 2。从表 2 可以看出专业版的句子数明显高于公众版，而两个版本的平均句子长度并无显著差异。

表 2 不同版本癌症综合信息句子基本统计信息对比

统计项	专业版	公众版	比率（专业版/公众版）
平均句子数	647.5	240.2	2.70
平均句子长度	22.44	23.01	0.98

有代表性的句法分析统计结果，见表 3。专业版和公众版的句法分析结果有显著差异。以平均句子长度比例 2.70 为基准，公众版中的专有名词（NNP）显著少于专业版，动词原形（VB）显著高于专业版。此外公众版中的限定词（DT）、动词的过去分词（VBN）、名词复数（NNS）和连词（CC）略高于专业版。考虑到 UMLS 术语多被句法分析识别为专有名词，该结果与术语统计结果相印证。前者说明专业版的专业术语（以专有名词 NNP 方式体现）显著高于公众版，而公众版更多使用动词的原形表达，例如 Uterine Sarcoma 的公众版中有如下描述：Being exposed to x – rays can increase the risk of uterine sarcoma，句中的 increase 的词性为动词（VN），这种表达较易理解。这与实验之前对数据的观察结果相符。经发现在公众版的一般信息版块，普遍会对癌症进行定义。而专业版则少有定义信息，为方便比对，选取同时具有癌症定义信息的骨髓增生异常综合征（Myelodysplastic Syndrome，MDS）的专业版定义和公众版定义，对其进行句法分析。原文和句法分析结果，见表 4。专业版的专有名词（NNP）数量高于公众版，动词（VB）和名词复数（NNS）却低于公众版，这与表 3 的结果相印证。考虑到很多较常见的癌症，例如膀胱癌，在专业版中直接省略了其定义信息，可以推测公众版关于癌症的定义和解释信息是词性统计结果差异的重要原因。

表 3 不同版本癌症综合信息词性分析统计结果对比

统计项	专业版	公众版	比率（专业版/公众版）
NN（名词）	453 612	190 092	2.39
IN（介词）	229 276	93 937	2.44
DT（限定词）	115 394	79 642	1.45
NNS（名词复数）	134 449	67 241	2.00
JJ（形容词）	196 585	65 889	2.98
NNP（专有名词）	243 116	36 148	6.73
CC（连词）	69 600	33 665	2.07
VBN（过去分词）	58 373	30 798	1.90
VB（动词）	28 554	29 696	0.96

表 4 癌症定义信息词性分析对比示例

版本	原文	POS 结果	结果统计
专业版	The MDS are a collection of myeloid malignancies characterized by one or more peripheral blood cytopenias.	The/DT MDS/NNP are/VBP a/DT collection/NN of/IN myeloid/JJ malignancies/NNS characterized/VBN by/IN one/CD or/CC more/JJR peripheral/JJ blood/NN cytopenias/NNS. /	NNP：1 VB：0 DT：2 VBN：1 NNS：2 CC：1
公众版	Myelodysplastic syndromes are a group of cancers in which immature blood cells in the bone marrow do not mature or become healthy blood cells.	Myelodysplastic/JJ syndromes/NNS are/VBP a/DT group/NN of/IN cancers/NNS in/IN which/WDT immature/JJ blood/NN cells/NNS in/IN the/DT bone/NN marrow/NN do/VBP not/RB mature/VB or/CC become/VB healthy/JJ blood/NN cells/NNS. /	NNP：0 VB：2 DT：2 VBN：0 NNS：4 CC：1

总结基于词汇和基于句法的分析结果，PDQ 在面向不同对象的资源建设过程中，在专业版词汇中加入了更多的术语量，尤其是 UMLS 术语，而公众版的面向公众的术语量略高于专业版，且难度分数偏低；专业版的句子数偏高，平均句子长度与公众版无明显差异；专业版中的专有名词数显著高于公众版，动词数则显著低于公众版。说明 PDQ 的专业版到公众版的转化，一方面考虑到术语的表现形式，另一方面在句法上也做了一定调整。

4 结语

本研究以癌症综合信息库 PDQ 为研究对象，从语言学特征角度解析医学指南知识库在面向患者教育目标时，基于面向医师版的知识库作出的调整。基于 OAC CHV 词典和基于句法分析的统计分析对比结果表明，面向患者的公众版使用较多的消费者词汇且难度分数偏低，篇幅较专业版显著降低，专有名词量偏低且动词原形量偏高。说明从语言学角度，面向患者与面向医师的知识库具有显著的统计差异，可供国内面向患者的指南类知识库建设提供借鉴。结合前期基于结构的研究成果，发现在 PDQ 中存在很多类似含义却表达不一致的语句，这些平行语句可构成很好的转化示范，在未来的工作中，将进一步通过语义相似度比对，构建平行语料库，供知识库建设参考或更深层次的语言学分析，例如依存句法分析等。

参考文献

1 Lawler F H, Viviani N. Patient and Physician Perspectives Regarding Treatment of Diabetes: compliance with practice guidelines [J]. Journal of Family Practice, 1997, 44 (4): 369 - 373.

2 Lehnbom E C, McLachlan A J, Brien J E. A Qualitative Study of Swedes' Opinions about Shared Electronic Health Records [J]. Stud Health Technol Inform, 2013, (192): 3 - 7.

3 GMC - UK. Good Medical Practice [EB/OL]. [2016 - 12 - 25]. https://www.gmc - uk.org/guidance/good_medical_practice.asp.

4 PDQ. PDQ XML Specification Document [EB/OL]. [2016 - 12 - 25]. http://www.cancer.gov/cancertopics/pdq.

5 Manrow R E, Beckwith M, Johnson L E. NCI's Physician Data Query (PDQ®) Cancer Information Summaries: History, Editorial Processes, Influence, and Reach [J]. J Canc Educ, 2014, 29 (1): 198 - 205.

6 李楠，姚亮，吴琼芳，等. 2012 - 2013 年中国大陆期刊发表临床实践指南质量评价 [J]. 中国循证医学杂志，2015, 15 (3): 259 - 263.

7 林夏，杨克虎，陈耀龙，等. 中国临床实践指南的现状与思考 [J]. 中国循证医学杂志，2017, 17 (5): 497 - 500.

8 Peters P, Smith A, Funk Y, et al. Language, Terminology and the Readability of Online Cancer Information [J]. Medical Humanities, 2015, 42 (1): 36 - 41.

9 孙月萍，侯震，侯丽，等. 面向医师和患者的医学教育信息资源建设探讨 [J]. 中国医学教育技术，2017, 31 (6): 655 - 657.

10 CHV. Consumer Health Vocabulary Initiative [EB/OL]. [2011 - 5 - 13]. http://consumerhealthvocab.org/docs/README.pdf.

11 De Marneffe M C, Maccartney B, Manning C D. Generating Typed Dependency Parses from Phrase Structure Parses [C]. Proc of LREC, 2006: 449 - 454.

(上接第 29 页)

参考文献

1 赵禹. 利用虚拟化技术瘦身医院 IT 系统 [J]. 中国管理信息化，2017, 20 (5): 142 - 144.

2 李先锋，王凯芸，吕强，等. 三甲医院虚拟化技术的研究与实践 [J]. 中国医院，2012, 16 (2): 12 - 14.

3 张钧，於煌，王相峰，等. 基于 VMware 虚拟化技术的医院信息化系统的实现 [J]. 中华医院管理杂志，2013, 29 (2): 108 - 110.

4 宋好好. 云计算信息系统信息安全等级保护测评关键技术研究 [J]. 信息网络安全，2015, 15 (9): 167 - 169.

5 钱朝阳，陆明胜. 浅谈超融合基础架构 [J]. 数字技术与应用，2016, 34 (9): 216 - 217, 220.