• 医学信息研究 •

面向复杂决策和知识发现的医学知识不确定性计算方法*

杜 建

（北京大学健康医疗大数据国家研究院 北京 100191）

〔摘要〕 将知识/证据的不确定性测度和结构化知识图谱相结合，总结并提出计算医学知识不确定性的几种方法，包括量表、概率、信息熵、证据－评论网络等。提出对于高确定性的知识，可由机器做决策；对于低确定性的知识，要触发人机交互，必须由机器和医生（科学家）共同决策，以此提高知识驱动的决策支持效率。

〔关键词〕 不确定性；科学知识；概率；信息熵；证据－评论网络

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2022.07.006

Approaches on Measuring the Uncertainty of Medical Knowledge for Complex Decision Making and Knowledge Discovery *DU Jian, National Institute of Health Data Science, Peking University, Beijing 100191, China*

〔**Abstract**〕 The paper combines the textual uncertainty knowledge (evidence) with structured knowledge graph, and puts forward several methods to calculate the uncertainty level of medical knowledge, including scale, probability, information entropy, evidence comment network and so on. It is proposed that the decision can be made by the machine for the knowledge with higher certainty level. For such conditions where there is only evidence with lower certainty level, it is important to join machines and doctors (scientists) together for shared decision－making, so as to improve the efficiency of knowledge－driven decision support.

〔**Keywords**〕 uncertainty; scientific knowledge; probability; information entropy; evidence comment network

1 引入知识不确定性分析的原因

1.1 从数据到知识、决策的基本特征和转化路径

在数据－信息－知识－智慧（Data－Information－Knowledge－Wisdom，DIKW）模型中，从数据到智慧其价值越来越高，但可计算性越来越低。如何将隐藏在科学文本大数据中的知识再次进行数据化是实现知识可计算性的关键途径。从数据到信息和知识主要依赖信息学方法（如本体）和数据科学方法（如机器学习），而从知识到智慧要解决如何在不确定性的条件下做出最佳决策的问题，见图 1。

〔收稿日期〕 2022－02－21

〔作者简介〕 杜建，博士，副研究员，发表论文 80 余篇。

〔基金项目〕 国家自然科学基金面上项目“不确定性科学知识表示与计量的理论、方法与应用研究：以医学为例”（项目编号：72074006）；中国科协青年人才托举工程项目“医学知识结构化表示与智能化计算模型研究”（项目编号：2017QNRC001）。

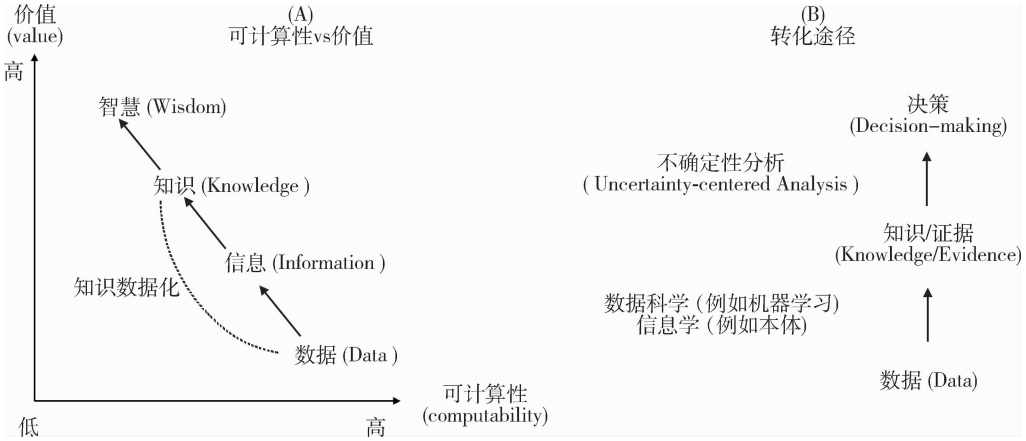

图 1 从数据到知识到决策的基本特征（A）和转化途径（B）

美国和欧洲正在进行描述可计算的生物医学知识（Computable Biomedical Knowledge，CBK）的元数据相关研究，即用哪些属性或字段描述可计算的医学知识。在 FAIR（Findable，Accessible，Interoperable，Reusable）原则基础上增加了 T（Trustable）[1]，强调确定性和可靠性在可计算的医学知识对于决策和应用中的重要性。例如增加对于证据基础的描述，包括知识数据来源、证据的确定性程度等。可见知识的确定性程度是知识表示不可或缺的要素。

1.2 复杂性问题的循证决策

1.2.1 *内涵与作用* 循证决策是借鉴循证医学而发展的一套决策理论，其认为政策和决策制定应吸收和使用最新科学证据，同时将社会经验和价值判断结合起来，做出最佳决策，尤其是重大突发公共卫生事件的防控和治疗决策。而在政策和实践中执行循证决策时需要克服的关键障碍在于知识缺口与不确定性，以及有争议、无关、相互矛盾的证据。将科学知识的不完备性和不确定性作为重点考虑因素能够降低制定不正确或非循证决策的风险[2]。不确定性作为科学知识的认知状态，尤其是科学探索中各种不确定、不完整和可能相互矛盾的信息，是监管机构针对新医疗措施进行风险评估和管理时参考的重要证据。

1.2.2 *启发式决策* 科学决策是在决策者的信息处理能力、时间和知识有限的情况下做出的。诺贝尔奖得主赫伯特·西蒙认为决策者的理性是有限的。启发式决策即有限理性的模型，是一种使用部分可用信息而忽略其余信息的决策策略，仅基于部分变量进行决策，不仅可以降低复杂度，而且可以提高决策的准确性、速度和透明度[3]。快速省力的启发式决策方法最初是在认知和决策科学的背景下提出的。医学、犯罪、商业、法律、体育等多学科领域的实践表明，这种决策方式得出的结论质量不亚于基于各种数据搜集、繁复计算得出的结论。普赖斯奖得主、德国马普学会文献计量学家 Lutz Bornmann 将其移植到科研评价决策过程中，提出基于文献计量学证据的启发式决策，例如将根据文献计量学的决策树模型（Bibliometrics - Based Decision Tree，BBDT）用于确定荷兰莱顿大学排名中的两所大学表现是否存在实质性差异[4]，为文献计量学方法广泛用于科研评价目的提供通用理论框架。基于此可以根据知识主张背后的科学证据的不确定性程度进行启发式决策，特别是涉及未知、不完备、不充分的科学认知和证据基础上进行决策时，使用启发式决策可能是最佳策略。

1.2.3 *循证启发式决策实施路径* 科学家与决策者之间沟通不畅、科学知识与决策过程脱节的原因之一在于决策者可能无法区分稳健、可信的科学证据与模糊、不确定的科学论断。所有科学都具有不确定性，只有有效地揭示和传递科学的不确定性才能更好地促进循证决策[5]。复杂网络领域的学者通过定量分析政策文件和科学论文之间的引用关系，揭示了科学和政策的共演化特征和循证决策模式，

发现很多政策文件引用了最新、经过同行评审、高影响力的科研成果[6]。提示政策界与学术界的紧密互动方式展现出二者之间的联系已经发挥了作用，但科学知识和政策之间具体、微观的交互模式尚不明确。政策和实践中的循证过程也是证据分析过程，是在一阶科学证据基础上进行二次分析，形成有助于决策的见解和知识。做好循证决策需要解决两个问题，一是跟上科学认知和证据的进展；二是解决早期信息缺失和中后期信息爆炸并存的问题。如果将科学证据的不确定性计算出来，或将具有矛盾性、冲突性的知识主张清晰揭示出来以辅助决策者进行参考，将有助于循证启发式决策。而决策总是在复杂现实中的多类因素相互动态作用下进行的，这种复杂性正是影响决策的关键问题，可借助成熟可靠的科学研究帮助解决[7]。

1.3 提高知识发现可靠性和效率

1.3.1 基于文献的知识发现 计算科学家从文献和数据库提取知识并进行计算处理，挖掘可以在实验中得到检验的新假设。实验科学家和计算科学家之间的合作已成为科学知识发现的新趋势[8]。这一概念与基于文献的知识发现是相似的，即融合零散、非相关的信息片段，揭示出有发展前景的新研究方向，或者提供潜在的变革性或突破性的见解[9]。基于文献的知识发现的最大需求、挑战、价值在于识别当前被忽视的研究领域，并结合其他信息识别未来值得科学界探索的前沿[10]。而科学研究前沿往往具有不确定性，特别是表现为未经验证的研究假说、冲突性、矛盾性的知识主张等。

1.3.2 元知识 元知识理论认为，从科学文献中挖掘知识不应仅关注知识本身，有关知识的知识即元知识也很重要，例如通过分析科学文本语境信息可评估特定命题在科学上的确定性程度[11]。科学知识具有客观和主观双重属性，要真正实现从现有知识大数据中再次发现新知识，不仅要关注结构化的知识单元，还要关注知识背景，即元知识[12]。与客观认识论相对应的是实践认识论，该观点对科学知识可以完全解释和编码的假设提出挑战，认为开发知识管理工具以及据此做出决策和判断需要考虑科学知识固有的模糊性、不确定性。而且科学知识是多维的，兼具抽象性与具体性、隐性与显性、集体性与个体性、发展性与静态性。认识到知识表达的多样性、模糊性、不确定性和不一致性才能更高效地发现新知识。将知识的动态性、不确定性、具象化和争议性等纳入计算过程，是确保知识发现有效性和可靠性的关键因素。

2 医学知识不确定性的类型

2.1 主要类型

美国国家癌症研究所（National Cancer Institute，NCI）对医学知识的不确定性进行分类，认为不确定性有 3 种来源或表现，即可能性（Probability）、模糊性（Ambiguity）、复杂性（Complexity）[13]。其中模糊性主要体现在对于结果的估计缺乏可靠性、可信性和充分性。复杂性这一类型不是由事件的不确定性（可能性）或缺乏可靠性、可信度或有关该事件的信息的充分性（模糊性）引起的，而是缘于事件和概念本身可能出现状态的多样性。以上 3 种知识不确定性的类型无法全部量化，见图 2。

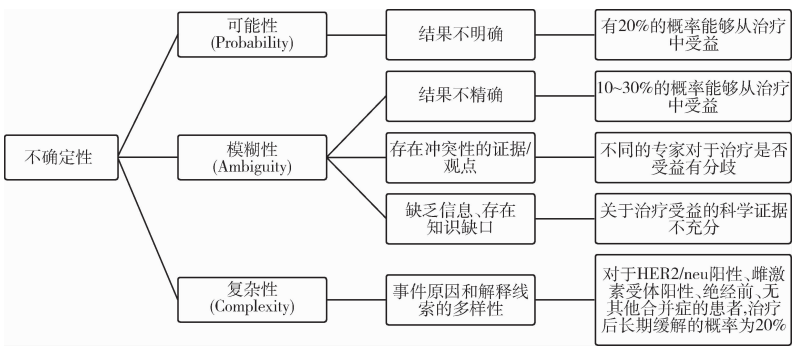

图 2 知识不确定性的 3 种类型（以乳腺癌治疗为例说明）

2.2 科学知识的不确定性程度能够反映科学发现的变革性程度

科学中假设推测的验证和争议矛盾的解决过程，分别对应渐进性研究和变革性研究。提示科学家发表研究成果时对科学发现表述的不确定性修辞和学术同行早期的争论式、批评式引用也是变革性的一类早期信号。科学文献遭遇负面引用并不总是说明该研究因无法重复而质量较低，需要分析负面引用在文献全文中的位置做出判断。结果和讨论部分的负面引用多缘于对数据结果的讨论，往往驱动在此基础之上开展进一步渐进性研究；而引言和结论部分的负面引用往往反映观点和概念分歧，更容易孕育变革性研究，其对科学前沿的预测意义更大。

2.3 不确定性科学知识表示与计算模型

科学知识主张主要通过科学出版物以文本形式表达，实现科学发现的可计算性应该深入到知识单元的微观层次，分析单元应侧重于观点和范式及其前提、证据和论证过程。因此提出面向知识发现、深入到知识单元和句子层面的不确定性科学知识表示与计算模型，其分为 4 个组件：编码；以三元组表示的知识单元；知识来源，即关于知识主张的陈述；认知状态，即不确定性分级（包括未知、假设推测、争议矛盾）[12]。进一步以该模型为基础，挖掘肺癌领域和心血管领域不确定性医学知识主张，尤其是争议性、冲突性、矛盾性的知识。该模型将以自然语言表达的海量知识主张进行结构化，并与其背后的数据或证据关联起来，既实现了细粒度表示知识对象的目标，又解决了当前知识发现研究忽略知识不确定性程度的问题，见图 3。

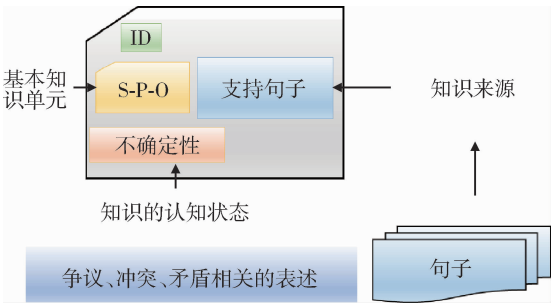

图 3 深入到知识单元和句子层面的不确定性科学知识表示模型

3 医学知识不确定性的表示和计算方法

3.1 量表

医学知识不确定性的表示和计算的核心在于将不确定性文本转化成数值、数字。如可以用量表、概率、证据评估、信息商等方法。美国学者将诊断报告的不确定性做成 Likert 量表[14]，根据诊断报告中表达不确定性的词和短语，例如“possible”“probable”“definite”“uncertain”“likely”“unlikely”“consistent with”“compatible with”“diagnostic of”“cannot exclude”医生和患者遵从上述体系建立标准，降低了信息丢失、信息误差，见图 4。

	−5	−3	−1	0	1	3	5
诊断的不确定性程度	非常不确定地排除一种疾病	有可能排除一种疾病	极有可能排除一种疾病	确诊	极有可能确诊一种疾病	有可能确诊一种疾病	非常不确定地提示着一种疾病
示例	检查方式可能有局限，但仍未见结节	未提示结节	未见明显结节	右上肺叶结节 无结节	疑似结节	提示结节	不能排除结节

图 4 诊断的不确定性示例

3.2 将修饰词转化为概率

荷兰学者从自然语言的完整语义表示的角度提出科学主张的形式化表示模型，该模型分为 4 部分：适用情境（Context class）、主语（Subject class）、修饰符（Qualifier）、关系（Relation type）；宾语（Object class）。为兼顾科学主张表达的机构化和完整性，在用“主语 - 关系 - 宾语”三元组表达核心知识主张的同时不能忽略科学主张的适用情境和修饰符[15]，例如在三元组中，“肥胖并伴有代谢异常 - 同时发生 - 膝盖骨关节炎”的适用情境是人。还有一个修饰词是 generally，对应着这一事件的发生概率，即可能性程度。可以将不同的线索词转化为概率，见表 1。这一词表可以拓展，如 likely、very likely 等。从长远来看，研究人员可以用上述模式表达其发现，从而将研究工作直接添加到科学发现的复杂知识图谱中。在此基础上可以开展查询类似研究、证实科学主张、发现矛盾、提供聚合和可视化、回答问题以及许多其他类型任务。

表 1 将表达不确定性的线索词转化为概率

修饰词	解释	概率值
(can) always	100%	=1
(can) generally	at least 90%	≥0.9
(can) mostly	at least 50%	≥0.5
(can) frequently	at least 10%	≥0.1
(can) sometimes	at least 0.1%	≥0.001
(can) never	0%	=0
(can) generally not	at most 10%	≤0.1
(can) mostly not	at most 50%	≤0.5
(can) frequently not	at most 90%	≤0.9
(can) sometimes not	at most 99.9%	≤0.999

3.3 证据评论的情感计算

本研究提出可以通过科学评论文本的情感计算方法对证据进行评估。科学评论是一类出版物，是指正式发表的短篇文章（例如观点、社论、评论、给编辑的信等），表达对所关注的原始研究支持性或反驳性的观点，或讨论其中的方法和发现，是对证据重要性和有效性进行科学评估的一种有效方式。以某种疾病药物治疗为例，早期关于新药治疗存在大量缺失、不确定、冲突甚至不准确的证据，通过 PubMed 获取被评论过的疾病相关文献作为证据（Evidence），以及疾病相关评论（Comment），构建证据 - 评论网络（Evidence - Comment Networks）。通过 PubTator 文本挖掘工具从标题/摘要句中抽取并识别常被评论的实体和概念。选择 6 组药物通过探索证据 - 评论网络的结构性和情感性信息，详细分析并重新生成经评论过滤后的证据主张。应用世界卫生组织（World Health Organization，WHO）指南对于这 6 类药物的使用建议作为金标准对照，以验证评论用于重塑临床知识主张的准确性、覆盖度和效率。分析结果表明，关于 6 类药物的证据被评论的积极/消极情感与 WHO 指南中对该药物使用的支持/反对建议完全一致。评论主题涵盖了证据评估的所有重要方面，以及方法学、临床适用性以外的其他方面，如伦理学、社会文化等。在时效性方面，50% 的批评性评论比指南发布时间平均提前了 4.25 个月。评论中还提示了表明临床实践中药物使用的不确定性，例如无法确定最佳剂量。笔者认为，评论可以作为一种快速证据评估工具，通过评估现有证据中的益处、局限性和其他临床实践问题而具有选择效应。科学评论可以帮助选择出重要的证据并对其有效性进行重塑。建议从信息学角度建立一个基于评论主题和情感取向的评分系统，以充分发挥科学评论在证据评估和不确定性决策中的潜力[16]。

3.4 信息熵

3.4.1 用信息熵测度知识不确定性程度 信息熵（Information Entropy，IE）概念是用于描述信源的不确定性。借鉴到医学的不确定性中，例如某条知识的表达是模糊、不完备甚至冲突、矛盾的，就发出了这样的信号。受陈超美相关研究启发[17]，提出用信息熵测度知识不确定性程度的方法。信息熵是反映事件不确定性的测量指标。其中事件即表示“模糊修饰”和“争议矛盾”的线索词是否出现。1 个知识单元（三元组）的不确定性，即信息熵 $U(t)$，等于与之相关的 n 个句子（$n \geq 1$）信息熵的总和：

$$U(t) = \sum_{1}^{n} U(s)$$

每个句子（sentence）的信息熵 $U(s)$，与该

句子中表示“模糊修饰”和“争议矛盾”的线索词（word）的概率 p（w）有关：

$$U(s) = -\sum_{(w \in s)} p(w) \cdot log(p(w))$$

这类词如果没有出现在句子中，则该句子的信息熵为 0，即该句子没有表达不确定性；这类词一旦出现在句子中，出现得越多则不确定性越高、信息熵越大。

这类词的概率 p（w）与该词在所有由句子表示的知识主张构成的知识体系中的出现频次有关。在医学领域中可以用 SemMedDB 中近 2 亿条能够抽取出三元组的句子中含每个词的句子数占总句子数的比例来计算。例如 2020 年最新版 SemMedDB 共含 214 721 135 个句子（PubMed 标题和摘要中的句子），其中“controvers *”（含 controversial 和 controversy）出现在 208 264 个句子中，该词在整个医学知识体系中的出现频率即概率是 0.000 969 91。通过计算表征不确定性的线索词在 SemMedDB 中的出现频次可知，所得信息熵的值与这些线索词的概率呈正相关。如 possible 的信息熵高于 controversial 是否一定说明用 possible 表达的知识比 controversial 表达的知识的不确定性程度要高，这一问题在科学机理上似乎难以解释清楚，见表 2。

表 2 表达假设推测和争议矛盾的线索词的概率及信息熵

线索词		在 SemMedDB 所有句子中的出现频次	信息熵（IE）
假设推测类	may/maybe	10 286	0.000 206 93
（Hedging lexicon）	possibl *	1 751 994	0.017 039 60
	potential	2 879 336	0.025 110 68
	seem *	333 677	0.004 364 49
	perhaps	84 058	0.001 333 87
	likely	1 052 986	0.011 325 48
	sometimes	119 942	0.001 817 05
矛盾争议类	conflict *	175 516	0.002 523 81
（Conflicting lexicon）	contradict *	46 639	0.000 795 66
	controvers *	208 264	0.002 922 65
	debat *	122 332	0.001 848 38
	disagree *	31 384	0.000 560 55
	disprov *	2 517	0.000 057 80
	no consensus	17 907	0.000 340 16
	questionable *	21 159	0.000 394 80
	refut *	9 710	0.000 196 47
	uncertain	227 014	0.003 146 19
	unknown	525 536	0.006 391 16

3.4.2 *科学知识认知状态不确定性的测度指标和方法* 单个线索词实际反映了认知状态。基于此提出科学知识认知状态不确定性的测度指标和方法。采用信息熵来测度认知状态的分布是离散还是集中。将认知状态作为变量 X，X 的取值总体上可以分为 4 类：未知的、不清楚；推测、假设；争议、矛盾、冲突；未明确表达不确定性。可以通过计算每个三元组的来源语句中 4 种状态的概率分布是集中还是离散来测度三元组认知状态的不确定性[18]。但在众多文献中，有一小部分知识主张（含表达不确定性的判断）是“原创的”，其余很多文献中的相关句子和判断其实是照搬效仿的，即受到了早期原创性论断的影响。如果能从时序上筛选出“原创的”、早期的主张，只分析这部分数据，可能得出的结论比“大数据”更可靠。今后拟继续研究这一问题。

3.4.3 *建立未知库* 科学通过“提出好问题”而进步，但生物医学文本挖掘相关研究尚未重点关注这些问题。在科学文献中发现科学问题或未知知识陈述不仅会产生新的文本挖掘工具，还会追踪学科中科学思想的演变，指出现有理论中的差距或缺陷，以及为未来洞察提供新途径[19]。因此相对于知识库提出建立未知库的构想。知识库主要包括先验知识，未知库则包括未知的知识，如尚未验证的科学假设、未解决的医疗问题或医疗需求。

4 结语

DIKW 模型中，从数据到信息和知识主要依赖信息学方法（如本体）和数据科学方法（如机器学习）；而从知识到智慧要解决的是如何在不确定性的条件下做出最佳决策的问题。将知识/证据的不确定性测度和结构化知识图谱相结合，为三元组配置置信度并提出置信度计算方法。对于高确定性的知识可由机器做决策；对于低确定性知识要触发人机交互，必须由机器和医生（科学家）一起做决策，以此提高知识驱动的决策支持效率。这也是将情报学与医学信息学进行交叉研究的一个方向。

参考文献

1 Alper B S, Flynn A, Bray B E, et al. Categorizing Metadata to Help Mobilize Computable Biomedical Knowledge [J]. Learn Health Syst, 2022, 6 (1): e10271.

2 Andermann A, Pang T, Newton J N, et al. Evidence for Health II: Overcoming Barriers to Using Evidence in Policy and Practice [J]. Health Res Policy Syst, 2016 (14): 17.

3 Bornmann L. Bibliometrics - Based Decision Trees (BBDTs) Based on Bibliometrics - Based Heuristics (BBHs): Visualized Guidelines for the Use of Bibliometrics in Research Evaluation [J]. Quantitative Science Studies, 2020 (1): 171 - 182.

4 Bornmann L, Marewski J N. Heuristics as Conceptual Lens for Understanding and Studying the Usage of Bibliometrics in Research Evaluation [J]. Scientometrics, 2019 (120): 419 - 459.

5 Fischhoff B, Davis A L. Communicating Scientific Uncertainty [J]. Proccedings of the National Academy of Sciences of the United States of America, 2014, 111 (Suppl 4): 13664 - 13671.

6 Yin Y, Gao J, Jones B F, et al. Coevolution of Policy and Science during the Pandemic [J]. Science, 2021, 371 (6525): 128 - 130.

7 张晓林. 从 Informetrics 到 Decision Intelligence: 呼唤知识发现研究的范式演变 [J]. 数据分析与知识发现, 2019 (1): 2.

8 Efstathiou S, Nydal R, Laegreid A, et al. Scientific Knowledge in the Age of Computation: Explicated, Computable and Manageable? [J]. THEORIA. An International Journal for Theory, History Foundations of Science, 2019, 34 (2): 213 - 236.

9 Smalheiser N R. Rediscovering Don Swanson: The Past, Present and Future of Literature - based Discovery [J]. Journal of Data and Information Science, 2017 (2): 43 - 64.

10 Smalheiser N R. Literature - based Discovery: Beyond the ABCs [J]. Journal of the American Society for Information Science and Technology, 2012, 63 (2): 218 - 224.

11 Evans J A, Foster J G. Metaknowledge [J]. Science, 2011, 331 (6018): 721 - 725.

12 Li X, Peng S, Du J. Towards Medical Knowmetrics: Representing and Computing Medical Knowledge Using Semantic Predications as the Knowledge Unit and the Uncertainty as the Knowledge Context [J]. Scientometrics, 2021, 126 (7): 6225 - 6251.

13 Han P K, Klein W M, Arora N K. Varieties of Uncertainty in Health Care: A Conceptual Taxonomy [J]. Med Decis Making, 2011, 31 (6): 828 - 838.

14 Reiner B I. Quantitative Analysis of Uncertainty in Medical Reporting: Creating a Standardized and Objective Methodology [J]. J Digit Imaging, 2018, 31 (2): 145 - 149.

15 Bucur C I, Kuhn J, Ceolin D, et al. Expressing High - level Scientific Claims with Formal Semantics [C]. USA: Proceedings of the 11th on Knowledge Capture Conference, 2021.

16 Wang S, Du J. A Comment - derived Evidence Appraisal Approach for Decision - making Using Uncertain Evidence [EB/OL]. [2021 - 12 - 21]. https://arxiv.org/abs/2112.02560.

17 Chen C. A Glimpse of the First Eight Months of the COVID - 19 Literature on Microsoft Academic Graph: Themes, Citation Contexts, and Uncertainties [J]. Front Res Metr Anal, 2020 (5): 607286.

18 Guo X, Chen Y C, Du J, et al. Extracting and Measuring Uncertain Biomedical Knowledge from Scientific Statement [J]. Journal of Data and Information Science, 2022, 7 (2): 6 - 30.

19 Boguslav M R, Salem N M, White E K, et al. Identifying and Classifying Goals for Scientific Knowledge [J]. Bioinformatics Advances, 2021, 1 (1): vbab012.