·循证决策·

运用德尔菲法构建公共卫生决策证据质量分级系统

商　雪[1,2,3*]　邓欣欣[1,2,3]　郭康乐[4]　周丽萱[1,2,3]　杨克虎[1,2,3]　李秀霞[1,2,3]
1. 兰州大学公共卫生学院　甘肃兰州　730000
2. 甘肃省循证医学与临床转化重点实验室　甘肃兰州　730000
3. 兰州大学基础医学院　甘肃兰州　730000
4. 甘肃省人民医院　甘肃兰州　730000

【摘　要】目的:构建适用于公共卫生决策领域的证据质量分级方法。方法:基于德尔菲调查法对 24 名专家进行 2 轮函询,形成证据评价的条目内容、评分细则及综合评价方法的共识。结果:2 轮函询后专家意见基本一致,问卷回收率均高于 80%,专家积极系数较好;权威系数均高于 0.85;Kendall 协调系数分别为 0.227 和 0.494,显示具有统计学意义($P<0.05$)。整合专家意见,最终构建 15 个一级条目、55 个二级条目,综合各部分条目评价结果(最高计 15 分),最终转化为高(>11)、中((8-11])、低((4-8])、极低(≤4)4 个级别的证据强度。结论:基于德尔菲调查法初步建立了公共卫生决策证据质量分级系统,具有较好的适用性和可行性,但其系统性能和可推广性有待进一步验证。

【关键词】GRADE 方法; 证据分级; 公共卫生; 德尔菲法
中图分类号:R197　文献标识码:A　doi:10.3969/j.issn.1674-2982.2023.10.010

Using the Delphi method to construct a system for grading the quality of evidence for public health decision-making

SHANG Xue[1,2,3], *DENG Xin-xin*[1,2,3], *GUO Kang-le*[4], *ZHOU Li-ying*[1,2,3], *YANG Ke-hu*[1,2,3], *LI Xiu-xia*[1,2,3]
1. School of Public Health, Lanzhou University, Lanzhou Gansu 730000, China
2. Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou Gansu 730000, China
3. School of Basic Medical Sciences, Lanzhou University, Lanzhou Gansu 730000, China
4. Gansu Provincial Hospital, Lanzhou Gansu 730000, China

【**Abstract**】Objective: To construct a grading method for evidence quality applicable to the field of public health decision-making. Methods: Based on Delphi survey method, 24 experts were consulted for two rounds of correspondence to form a consensus on the contents of evidence evaluation items, scoring rules and comprehensive evaluation methods. Results: With the recovery rate being higher than 80% in the two rounds of correspondence interview, the experts' opinions were seen basically the same, and their positive coefficient was good. The authority coefficient was higher than 0.85. Kendall coordination coefficients were 0.227 and 0.494, respectively, showing statistical significance ($P<0.05$). According to the experts' opinion, 15 first-level items, 55 second-level items and 25 third-level items were finally constructed, and the evidence strength was finally translated into four levels based on the evaluation results of each part of items (up to 15 points): high (>11), moderate ((8-11]), low ((4-8]) and very low (≤4). Conclusions: Based on Delphi survey method, the grading system for evidence quality in public health decision-making is initially established, which has good applicability and feasibility, but its system performance and extensibility need to be further verified.

【**Key words**】GRADE; Grading of evidence; Public health; Delphi method

* 基金项目:国家自然科学基金项目(72074103)
作者简介:商雪(1998 年—),女,硕士研究生,主要研究方向为循证卫生政策与管理。E-mail:1091844882@qq.com
通讯作者:李秀霞。E-mail:lixiuxia@lzu.edu.cn

1 背景

制定公共卫生建议或政策依赖于对一系列复杂因素的判断,包括卫生问题的严重程度、特定干预的益处和危害、人力和财政资源的使用、可转移性以及干预的可接受性和可行性。[1] 公共卫生的干预措施通常影响到广泛的人群,可以在个人和人群层面产生重大的健康效益。[2-3] 关于如何收集、综合公共卫生证据并将其用于决策还存在很多讨论,使决策过程明确和透明,仔细审查所依据的证据类型至关重要。不同国家的公共卫生组织制定了不同的方法来评估证据的质量[4-6],同时使用许多不同的方案可能会导致对证据质量的不同评级和相互冲突的建议,这可能会阻碍指南制定者和决策者以透明的方式做出明智决策的目标[7]。

现阶段,大部分证据分级系统中均关注到了研究设计、研究的偏倚风险、研究质量和研究局限性、精确性、直接性和一致性等方面,但在具体的评定和标准设置上仍存在诸多的不适用。[8] 以往系统中忽视了公共卫生的背景特征,如健康分布(社会不平等指标)、健康决定因素(因果网络)、健康对个人和社会的影响和改变健康决定因素的方法[9]、证据使用中的一系列促进和阻碍因素等。可能难以反映证据的真实质量,从而影响证据的转化,缺乏对公共卫生决策进行证据质量分级的特色理论。即使是广泛使用的 GRADE 工具在公共卫生领域应用中也存在诸多挑战。例如,不同类型的观察性研究起始证据等级低且相同,难以反映该领域真实证据质量;主观性较大,缺乏具有明确阈值的定量评价方法等。[10]

鉴于目前我国尚缺乏适用于公共卫生复杂环境的分级体系,且现有的分级体系构建较早,尚不完善,缺乏针对性,故本研究充分考虑公共卫生领域的需求,邀请国内外专家开展德尔菲调查,基于专家共识,构建适用性、科学性、可操作强的公共卫生决策证据质量评价体系(Evidence quality grading system for public health decison-making, PHE-Grading),以提升公共卫生证据质量,促进政府及卫生人员合理决策。

2 资料与方法

2.1 资料来源

在方法学专家指导下,课题组通过检索中国知网、万方、中文科技期刊数据库(VIP)、中国生物医学文献数据库(CBM)、PubMed、WOS、Cochrane Library、EmBase 等数据库,并参考教材、专著、指南、标准、共识、规范等,收集公共卫生证据质量分级相关的评价条目,基于主题综合法整理、编码、罗列所有条目,构建备选条目池。

2.2 德尔菲调查法

2.2.1 专家遴选

本研究通过网络查询、电话咨询、文献报告及行业专家推荐获得专家信息,考虑多学科交叉和公共卫生领域的特点,在国内外遴选出 24 名专家进行德尔菲函询调查。

2.2.2 问卷指标的评价方法

按照初步构建的条目框架编制专家函询问卷,采用 Likert 5 级评分法对两轮专家选取的公共卫生决策证据质量分级方法条目进行评价,每个条目按照重要性设置 5 个等级,选项为 1 ~ 5 分,未作答条目视为不确定。

2.2.3 专家函询

通过电子邮件(E-mail)和短信方式向专家发放函询问卷(问卷星)。邀请 5 名专家进行预函询,结果显示,专家函询问卷内容效度指数(CVI)为 0.848,表明该问卷具有良好的整体效度。根据专家意见调整函询内容,最终制订出包括 17 项一级条目、59 项二级条目的第一轮专家函询问卷。第二轮专家咨询问卷在第一轮专家咨询的基础上根据专家意见进行修改、整理,邀请专家进行第二轮评价。专家函询拟开展 2 ~ 3 轮调查,直至专家意见形成一致。[11]

2.2.4 条目筛选标准

根据指标集中程度和变异程度进行筛选。专家函询条目中,条目筛选标准为:(1) 条目满分率 > 40%;(2) 重要性均值 ≥ 4 分;(3) 等级和 > 70%;(4) 变异系数 < 0.25。[12] 基于统计分析结果,若满足以上 4 项标准中的 3 项及以上,则条目予以保留,满足 2 项则在第二轮咨询中进一步讨论,若满足 1 项或均不满足则删除。

2.3 统计分析

通过 Excel 软件进行数据整理与分析,计算构成比、均数、标准差及变异系数。应用 SPSS 22.0 统计分析专家积极系数、专家意见集中程度、专家权威系数及专家意见协调程度。[13] (1) 专家积极系数:以专家问卷有效回收率和专家意见提出率表示[14],回收率 ≥ 75% 可确保问卷的准确性。(2) 权威程度系数:权威程度系数(Cr)由专家对条目评价的判断依据(Ca)和熟悉程度(Cs)决定,即 Cr = (Ca + Cs)/2,

一般认为 Cr > 0.7 即可接受。[15] 专家对条目的判断依据分为实践经验、理论分析、参考国内外文献、直观感受 4 个维度，每个维度划分为大、中、小不同程度；不同层次赋分为：实践经验（0.5、0.4、0.3）、理论分析（0.3、0.2、0.1）、同行了解（0.1、0.1、0.1）、专家直觉（0.1、0.1、0.1）。专家对所咨询条目的熟悉程度分为：很熟悉（1.0）、比较熟悉（0.8）、一般熟悉（0.6）、不大熟悉（0.4）和不熟悉（0.2）5 个层次。[16]（3）专家意见集中程度：计算各条目的均数（$\overline{X}$）、等级和（S）、满分率（K），其分值越大，重要性越高。（4）专家意见协调程度：计算变异系数（CV）及 Kendall's W 协调系数（W）。变异系数大于 0.25，表明专家分歧较大。协调系数在 0 ~ 1 之间，越接近 1 表明专家协调程度越高，采用 χ^2 检验进行分析。[17]

3 结果

3.1 专家的基本情况

如表 1 所示，本研究共遴选国内外专家 24 名。年龄集中在 31 ~ 40 岁；文化程度均为硕士和博士；专家大部分从事于多个研究领域，其中循证医学领域专家最多（66.67%）；主要来自高校/科研机构（70.83%）；正高级职称占比 58.33%；工作年限在 5 ~ 10 年占比最高（45.83%）。大约有 29.16% 的专家非常熟悉证据分级系统，其中最为熟知的为 GRADE 证据分级系统；主要将证据工具应用于临床实践指南（表 2）。

表 1　专家基本信息

	数量	构成比（%）
性别		
男	14	58.33
女	10	41.67
年龄（岁）		
0 ~	3	12.50
31 ~	13	54.17
41 ~	4	16.67
51 ~	2	8.33
61 ~	2	8.33
工作单位类别		
高校/科研机构	17	70.83
医疗机构	7	29.17
职称		
正高级	14	58.33
副高级	7	29.17
初级	3	12.50
工作年限（年）		
0 ~	4	16.67
5 ~	11	45.83
11 ~	5	20.83
21 ~	1	4.17
31 ~	3	12.50
文化程度		
硕士	4	16.67
博士	20	83.33
专业		
循证医学	16	66.67
公共卫生与预防医学	11	45.83
证据评价及决策转化	7	29.17
多学科交叉领域	7	29.17
医院管理与卫生政策	5	20.83
中医/中西医结合/中药学	3	12.50
循证药学	1	4.17
方法学/数据挖掘	2	8.33
地域分布		
加拿大	2	8.33
瑞典	1	4.17
北京	4	16.66
甘肃	6	25.00
上海	2	8.33
四川	5	20.83
山东	1	4.17
江苏	1	4.17
陕西	1	4.17
贵州	1	4.17

表 2　专家对分级系统的熟悉程度

类别	数量	构成比（%）
您了解证据分级系统吗？		
非常了解	7	29.16
了解	17	70.84
了解较少	0	0.00
不了解	0	0.00
您熟悉的证据分级系统？		
加拿大定期健康体检工作组分级	6	25.00
英格兰北部循证指南制定证据分级	3	12.50
苏格兰院际指南网络分级	7	29.16
证据金字塔	8	33.33
英国牛津循证医学中心证据分级	12	54.17
WHO 分级标准	11	45.83
中国循证医学中心证据分级	5	20.83
美国卫生保健政策研究所分级	5	20.83
GRADE 分级	24	100.00
NutriGrade 系统	10	41.67
您主要将证据分级方法应用于那些领域？		
卫生技术评估	9	37.50
临床决策分析	11	45.83
公共卫生决策	11	45.83
临床实践指南	20	83.33
系统评价/Meta 分析	20	83.33
网状 Meta 分析	10	41.67
诊断性试验	5	20.83

3.2 专家的积极性和权威程度

对24名专家发放两轮函询问卷，第一轮应答率为100.00%，第二轮应答率83.33%。两轮函询专家的意见提出率分别为75.00%和35.00%，专家的积极系数较好。两轮函询专家权威系数分别为0.86和0.87，专家对条目有较高的把握，咨询结果可靠性好。

3.3 专家意见的集中程度

如表3所示，在第一轮函询中，有12个条目平均分大于等于4分；满分比大于40%的有10条；等级和S大于70的条目有17条，这些条目在下一轮咨询中十分重要。第二轮函询结果显示（表4），每项条目赋值的算术均数为2.55～4.50，满分率为10.00%～75.00%，等级和为51～90，各专家评价意见比较一致。

表3　第一轮关键条目遴选的专家咨询结果

分级条目	最低分	最高分	$\overline{X} \pm S$	例数	K(%)	S	CV
研究设计	4	5	4.71±0.46	24	70.83	113	0.10
研究执行质量	4	5	4.83±0.38	24	83.33	116	0.08
一致性	2	5	4.50±0.78	24	66.67	108	0.17
精确性	3	5	4.33±0.64	24	41.67	104	0.15
直接性	3	5	4.29±0.69	24	41.67	103	0.16
发表偏倚	2	5	4.00±0.83	24	25.00	96	0.21
大效应量	3	5	4.33±0.70	24	45.83	104	0.16
剂量—反应关系	2	5	4.25±0.85	24	45.83	102	0.20
负偏倚	3	5	4.25±0.68	24	42.50	90	0.18
稳健性	2	5	3.88±0.99	24	29.17	93	0.26
因果推断重要性	2	5	4.33±0.87	24	50.00	104	0.20
资助偏倚	2	5	4.00±0.89	24	29.17	96	0.22
资源依赖度	1	5	3.54±1.06	24	16.67	85	0.30
可推广性	1	5	4.08±1.10	24	45.83	98	0.27
利弊平衡	1	5	3.67±1.13	24	29.17	88	0.31
健康公平性	1	5	3.58±1.10	24	20.83	86	0.31
证据的阻碍和促进因素	1	5	3.79±1.06	24	33.33	91	0.28

注：$\overline{X} \pm S$表示均数加减标准差，K(%)表示满分比，S表示等级和，CV表示变异系数。

表4　第二轮关键条目遴选的专家咨询结果

分级条目	最低分	最高分	$\overline{X} \pm S$	例数	K(%)	S	CV
稳健性	3	5	4.35±0.67	20	45.00	87	0.15
资源依赖度	1	5	2.55±1.23	20	15.00	51	0.48
利弊平衡	1	5	3.50±0.89	20	10.00	70	0.25
健康公平性	1	5	4.20±0.89	20	35.00	84	0.21
证据的阻碍和促进因素	1	5	4.50±1.05	20	75.00	90	0.23

注：$\overline{X} \pm S$表示均数加减标准差，K(%)表示满分比，S表示等级和，CV表示变异系数。

3.4 专家意见的协调程度

第一轮函询的变异系数为0.08～0.31，第二轮函询的变异系数为0.15～0.48；卡方分析发现第一轮函询 $W=0.227$，$\chi^2=83.04$，$P<0.001$，专家对各条目评价维度的意见基本一致。第二轮函询 $W=0.494$，$\chi^2=39.49$，$P<0.05$，专家对条目的认同度高，评分意见具有一致性。

3.5 专家咨询意见汇总结果

第一轮函询问卷包括17项一级条目、59项二级条目分级体系，经函询后，18名专家共提出48条意见，根据专家意见，作如下修改：（1）专家意见集中保留11项一级条目，建议删除稳健性、资源依赖度、利弊平衡、健康公平性及证据的阻碍和促进因素5个条目，但课题组认为这几个条目在公共卫生证据评价中具有意义，暂时予以保留，纳入第二轮函询中重

点讨论。(2)建议删除两项二级条目:“5.2 与 PICO 标准大致相符(即研究的人群、干预、比较和结局,3 项一致);6.1.4 纳入研究数 <5 篇”。(3)建议新增三项二级条目:“1.11 间断时间序列分析、回归不连续性设计;3.3 无法合成的定性研究;4.3 单个研究、定性研究”。(4)根据专家意见,修改 18 项二级条目:专家提出高和低偏倚风险的评定可能会混淆“几乎所有低,至少一个高”,故将低偏倚风险的评定调整为“2.1 若所有的纳入研究均被判定为低偏倚风险”;在效能一致性的评估中将 PICO 定义的异质性补充在内;专家指出精确性的衡量无法统一,具体数值的界定应谨慎,故对精确性的阈值设定作了相应调整;将直接性的分级合并为三类;在大中小效应中列举了实例;在资助偏倚中融入利益声明,在“私人机构、基金会、非政府组织赞助”部分区分了盈利与否的组织特性,及“无基金资助”的正向和负向性;在剂量—反应中补充了“不需要考虑剂量反应关系”的情况。(5) 证据的总体等级评定显示协调程度较差(表 5),专家建议需要考虑明确 2/4 和 3/4 所包括的区间,根据意见将 2/4 划分在低质量等级,3/4 划分在中等质量等级。

课题组对第一轮函询结果进行讨论,经调整后开展第二轮函询,函询结束后,经课题组协商,对条目体系做以下调整:(1)删除两项一级条目:“资源依赖度”和“利弊平衡”。(2)修改部分条目的语言表述:调整表达顺序 “3.1 一致性较好(如,PICO 一致,效应量的大小和方向一致,可信区间重叠度高,$I^2 \leq 50\%$;3.2 一致性较差(如,PICO 不一致,效应量的大小和方向不一致,置信区间无重叠;$I^2 > 50$)”;将“权重最小(即精确性最低)”修改为“具有严重偏倚”。(3)增加两个条目注释:其一,“精确性的评定需要根据具体的研究问题进行评价,不同的研究界值是否相同,需结合实际情况,明确目标人群在现实世界中占有怎样的比例,然后根据此类人群的占比情况设定相应的样本量阈值,进行灵活应对和解释。”其二,“评估效应量大小时应结合问题的严重程度,并与样本量对照分析”。(4)证据的总体等级评定一致性较好,划分为高、中、低和极低四个等级(表 5)。

表 5　证据总体等级评定的专家咨询结果

总体等级评定	例数	均值	标准差	变异系数
第一轮				
若评价得分低于条目总分的 1/4 则评定为极低质量	24	4.13	1.19	0.29
若评价得分在条目总分的 1/4 ~2/4 间则评定为低质量	24	4.25	0.85	0.20
若评价得分在条目总分的 2/4 ~3/4 间则评定为中等质量	24	4.38	0.77	0.18
若评价得分高于条目总分的 3/4 则评定为高质量	24	4.54	0.59	0.13
第二轮				
极低质量(评价得分小于等于条目总分的 1/4)	4.90	20	0.31	0.06
低质量(评价得分在条目总分的(1/4-2/4] 间)	4.90	20	0.31	0.06
中等质量(评价得分在条目总分的(2/4-3/4] 间)	4.90	20	0.31	0.06
高质量(评价得分大于条目总分的 3/4)	4.90	20	0.31	0.06

3.6 公共卫生决策证据质量分级系统

经两轮函询和课题组讨论补充修改后,最终构建公共卫生决策证据质量分级方法条目评价体系由 15 个一级条目、55 个二级条目构成(表 6)。综合各部分条目的评估结果(总体评分最高计 15 分),最终转化为高(>11)、中((8-11])、低((4-8])、极低(≤4)4 个级别的证据强度(表 7)。

4 讨论

4.1 公共卫生决策证据质量分级体系构建的重要性

随着我国公共卫生政策的不断推进,证据评价体系需要不断更新完善。目前国外已研发了一些系统应用于公共卫生项目,如美国预防服务工作队、加拿大预防保健工作队和英国国家健康与临床卓越研究所等,应用最为广泛的为 GRADE 分级系统。鉴于卫生政策、卫生系统或环境卫生干预往往比临床、筛查干预更复杂多样,尚没有统一的标准,完全照搬 GRADE 应用于公共卫生领域,往往无法很好地凸显公共卫生的研究特点及证据优势。因此,基于公共卫生背景研制适用、可行、可及的公共卫生证据质量评价系统对研究者及卫生决策者具有理论指导和实践双重意义。

表 6　公共卫生决策证据质量分级系统

一级条目（最高计分）	二级条目（计分）	一级条目（最高计分）	二级条目（计分）
1. 研究设计（1 分）	1.1 随机对照试验(1 分) 1.2 半随机对照试验(1 分) 1.3 队列研究(1 分) 1.4 病例—对照研究(0.5 分) 1.5 横断面研究(0.5 分) 1.6 病例报告/系列分析(0.5 分) 1.7 自身前后对照研究(0.5 分) 1.8 质性研究(0.5 分) 1.9 专家意见(0.5 分) 1.10 官方标准(0.5 分) 1.11 间断时间序列分析、回归不连续性设计(0.5 分)	7. 大效应量（1 分）	7.1 效应量非常大(1 分) 7.2 大效应量(1 分) 7.3 中等效应量(0.5 分) 7.4 无效应(0 分)
2. 研究执行质量(1 分)	2.1 低偏倚风险(1 分) 2.2 高偏倚风险(0 分) 2.3 不清楚的偏倚风险(0.5 分)	8. 剂量—反应关系(1 分)	8.1 无显著剂量—反应关系 (0 分) 8.2 存在剂量—反应关系(1 分) 8.3 未进行或不需要考虑剂量—反应关系(0.5 分)
3. 一致性（1 分）	3.1 一致性较好(1 分) 3.2 一致性较差(0 分) 3.3 单个研究、定性研究(0.5 分)	9. 负偏倚（1 分）	9.1 有个别研究存在负偏倚(0.5 分) 9.2 有多个研究存在负偏倚(1 分) 9.3 无研究存在负偏倚(0 分)
4. 精确性（1 分）	4.1 精确性较差(0 分) 4.2 精确性较好(1 分) 4.3 定性研究(0.5 分)	10. 稳健性（1 分）	10.1 稳健性良好(1 分) 10.2 稳健性较差(0 分)
5. 直接性（1 分）	5.1 与 PICO 标准基本相符(0.5 分) 5.2 与 PICO 标准完全相符(1 分) 5.3 与 PICO 标准不相符(0 分) 5.4 间接比较证据(0.5 分)	11. 因果推断重要性(1 分)	11.1 因果推断意义重大(1 分) 11.2 因果推断意义较小(0 分)
6. 发表偏倚（1 分）	6.1 严重的发表偏倚(0 分) 6.2 不清楚的发表偏倚(0.5 分) 6.3 中度的发表偏倚(0.5 分) 6.4 无发表偏倚(1 分) 6.5 未检测发表偏倚(0.5 分)	12. 资助偏倚（1 分）	12.1 行业资助,未进行利益声明或存在利益冲突(0 分) 12.2 私人机构、基金会、非政府组织赞助(非盈利性)(0.5 分) 12.3 学术机构、研究机构赞助(1 分) 12.4 无基金资助,声明无利益冲突(1 分)
		13. 可推广性（1 分）	13.1 推广性较好(1 分) 13.2 推广性较差(0 分)
		14. 健康公平性(1 分)	14.1 证据充分考虑了健康公平性(1 分) 14.2 仅考虑部分健康公平性(0.5 分) 14.3 未考虑健康公平性(0 分)
		15. 证据的阻碍和促进因素(1 分)	15.1 列出充分的证据使用促进和阻碍因素(1 分) 15.2 列出不充分的证据使用促进和阻碍因素(0.5 分) 15.3 未列出证据使用的促进和阻碍因素(0 分)

表 7　证据等级评定

评分	证据分级	具体描述
≤4	极低	对效应估计值几乎没有信心:真实值很可能与估计值大不相同
(4-8]	低	对效应估计值的确信程度有限:真实值可能与估计值大不相同
(8-11]	中	对效应估计值有中等程度的信心:真实值有可能接近估计值,但仍存在二者大不相同的可能性
>11	高	非常确信真实的效应值接近效应估计值

注:由于文本篇幅限制,无法呈现全部内容,如有需要,可联系作者获取。

本研究针对公共卫生领域证据分级工具应用存在的问题,在对公共卫生证据质量评价体系进行系统评价的基础上,从专家角度对纳入的分级条目进行遴选,最终形成共识构建证据质量分级系统。基于权威专家共识的公共卫生证据分级系统构建更为可靠、科学,有助于决策者恰当使用证据,推动公共卫生领域健康发展,扩展和丰富证据体系的应用,对未来研究的开展亦具有参考意义。

4.2 公共卫生决策证据质量分级体系构建的科学性

本研究主要综合了系统评价和德尔菲函询方法,进行条目筛选及体系构建。通过对公共卫生专家进行德尔菲调查,以相关性、可靠性、适用性、有效性作为主要筛选标准,形成了公共卫生决策证据质量分级工具,具有较好的实用性和可行性。数据分析显示专家的积极系数较好、权威程度高、协调程度较好,条目评价意见集中、具有较好的一致性,条目体系和水平层级具有高可信度。

4.3 公共卫生决策证据质量分级体系构建的创新与应用

经过德尔菲调查后,最终构建了由 15 个一级条

目、55 个二级条目组成的公共卫生决策证据质量分级工具。采用评分制而非升降级制，条目总分共 15 分，弥补了 GRADE 工具缺乏量化系统的不足。在本体系中，有 9 个条目与 GRADE 一致，但具体评分细则有所不同。

①研究设计，不作初始评级评定，而是根据不同研究类型赋予分值，可应对公共卫生领域证据水平分布不均，RCT 研究证据稀缺，不同类型的观察性研究起始证据等级均低的问题；②研究的执行质量，旨在评估构成证据体系的不同研究设计的研究执行情况，以尽量减少对内部和外部有效性的威胁，根据偏倚风险的高低进行判定；③一致性，指效应估计的大小和/或方向的相似程度，主要根据效应量的大小和方向，可信区间重叠度，I^2 值大小协助评分；④精确性，指对研究结果效应量估计值的把握度，评定方法与 GRADE 标准一致，其具体阈值的界定较复杂，需根据问题的成熟度，结合实际情况，进行灵活应对和解释。另外，本系统在一致性和精确性评定中考虑了单个和定性研究无法合并及量化的情况。⑤直接性，旨在评估目标问题和纳入研究之间的相似性，结合 PICO 标准进行判断；⑥发表偏倚，旨在评估研究结果的性质和方向导致研究成果的发表与未发表引起的偏倚，结合漏斗图和统计检验进行评分；⑦大效应量，即当方法学严谨的观察性研究显示疗效显著或非常显著且结果高度一致时，该条目的评估需结合问题的背景与样本量对照分析；⑨负偏倚，指当影响观察性研究的偏倚不是夸大，而可能是低估效果时，可提高其证据质量，基于负偏倚数量进行评分。

新增稳健性、因果推断重要性、资助偏倚、可推广性、健康公平性及证据的阻碍和促进因素 6 个分级条目。新增条目中，证据的稳健性十分重要，主要以敏感性分析进行判定。因果推断重要性条目的加入为评估不同类型证据与复杂干预措施之间的关系提供了重要框架。同时，引入了资助偏倚条目，所有资金，无论来自公共或私人，政府或行业资助都有可能对研究结果产生潜在的重要影响，尤其行业资助，会进一步降低证据的可信度。另外，证据的可推广性也被作为评估条目，用以检验研究结果是否普遍适用于特定研究范围之外的其它时间、情境和人群。GRADE 方法没有考虑公共卫生的背景特征，本系统将健康公平性作为评估条目考虑在内，确保公众都能公平享有健康服务。最后，证据使用中存在的促进和阻碍因素也被作为公共卫生决策研究领域评估的重要条目。最后，本工具以 GRADE 证据水平评价原则为主体，通过各条目评分的高低，分为高、中、低和极低 4 个等级。

本分级体系是基于 GRADE 衍生的一种更适用于公共卫生决策研究领域的证据质量分级系统，该系统既兼具 GRADE 方法的特色，又充分考虑了公共卫生决策领域研究的特定要求，基于不同研究设计类型，引入量化系统，重点考虑了公共卫生领域证据评价存在的局限性，整体评价过程客观透明、科学严谨，实用性和可行性强，大致需要 15 分钟（时间范围：10～30 分钟）。因此，在公共卫生决策领域，推荐使用该分级系统。

4.4 相比于已发表研究

Orton 的研究综合了关于公共卫生决策者在全民卫生保健系统环境中使用研究证据的经验证据，充分描述了研究证据使用的障碍，研究发现不同环境下的决策过程差异很大，关键参与者的看法也不同。[18] Schwingshackl 研发的 Nutri-grade 证据质量评分系统，其只适用于营养领域，范围较小。[19] Movsisyan 的研究系统审查了卫生和社会政策干预有效性证据质量评级系统，确定了包括研究设计、研究执行、一致性、精确性等在内的 13 个证据域，研究发现这些证据域的评级标准存在重大差异。[20] 相比于上述研究，本研究充分考虑了上述影响因素，制定了更加全面的评价标准和分级系统。

4.5 局限性

作为新衍生的分级系统，存在一定局限性，如：由于公共卫生证据分级系统条目构成及方法评估的多样性及复杂性，不同条目可能有相对重要性，且难以证明条目权重分配的科学性，目前尚未对各分级条目的评分值赋予权重，后续研究中将作进一步探索；尽管本研究引入公共卫生领域的 Meta 分析进行应用评价，但其系统性能仍待挖掘，未来仍需继续关注该分级系统在实际应用中性能的报告结果，以达到更好的推广应用。

5 结论

基于德尔菲函询构建的公共卫生决策证据质量分级系统，具有一定的合理性和适用性。鉴于该系统初步构建，尚需进一步开展实证研究予以验证其可执行性和可操作性。建议更多的方法学家、学术专家、研究人员和使用者结合其它评价体系使用，以

作更多的探讨、优化和评价，制定更全面的评价标准，构建更加科学可行、适用性、可操作性强的公共卫生证据分级工具，推进公共卫生事业不断发展。

作者声明本文无实际或潜在的利益冲突。

参 考 文 献

[1] Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance [J]. Int J Nurs Stud, 2013, 50(5): 587-592.

[2] Xun Y, Guo Q, Ren M, et al. Characteristics of the sources, evaluation, and grading of the certainty of evidence in systematic reviews in public health: A methodological study [J]. Front Public Health, 2023, 11: 998588.

[3] Rehfuess E A, Akl E A. Current experience with applying the GRADE approach to public health interventions: an empirical study [J]. BMC public health, 2013, 13: 9.

[4] Kelly M P. The need for a rationalist turn in evidence-based medicine [J]. Journal of evaluation in clinical practice, 2018, 24(5): 1158-1165.

[5] Vogt R L, Heck P R, Mestechkin R M, et al. Experiment aversion among clinicians and the public - an obstacle to evidence-based medicine and public health[J]. medRxiv, 2023(preprint).

[6] Hunter D J. Relationship between evidence and policy: a case of evidence-based policy or policy-based evidence? [J]. Public health, 2009, 123(9): 583-586.

[7] Levay P, Heath A, Tuvey D. Efficient searching for NICE public health guidelines: Would using fewer sources still find the evidence? [J]. Res Synth Methods, 2022, 13(6): 760-789.

[8] Tugwell P, DE Savigny D, Hawker G, et al. Applying clinical epidemiological methods to health equity: the equity effectiveness loop [J]. BMJ, 2006, 332(7537): 358-361.

[9] Practice guideline update recommendations summary: Disorders of consciousness: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National Institute on Disability, Independent Living, and Rehabilitation Research [J]. Neurology, 2019, 93(3): 135.

[10] Tomlinson E, Pardo J, Sivesind T, et al. Prioritising Cochrane reviews to be updated with health equity focus [J]. Int J Equity Health, 2023, 22(1): 81.

[11] Baicker K, Chandra A. Evidence-Based Health Policy [J]. The New England journal of medicine, 2017, 377(25): 2413-2415.

[12] Guyatt G H, Schünemann H J, Djulbegovic B, et al. Guideline panels should not GRADE good practice statements [J]. Journal of clinical epidemiology, 2015, 68(5): 597-600.

[13] 郑忠礼，宋旭萍，赵琳，等. 中医外治法治疗近视的系统评价再评价 [J]. 兰州大学学报(医学版)，2022，48(5)：79-86.

[14] 赵倩. 中医护理发展面临问题分析及策略思考 [J]. 健康之路，2017，16(5)：238.

[15] 王春枝，斯琴. 德尔菲法中的数据统计处理方法及其应用研究 [J]. 内蒙古财经学院学报(综合版)，2011，9(4)：92-96.

[16] 夏萍，吴大嵘，卢传坚，等. Delphi 法在医疗质量评价指标体系中的可靠性分析 [J]. 现代预防医学，2012，39(14)：3488-3490.

[17] 王珏莲，侯政昆，潘静琳，等. 基于德尔菲法的胃食管反流病(食管癔/吐酸)医生报告结局量表的研制与条目筛选 [J]. 中国中西医结合消化杂志，2019，27(10)：748-52.

[18] Orton L, Lloyd-williams F, Taylor-robinson D, et al. The use of research evidence in public health decision making processes: systematic review [J]. PloS one, 2011, 6(7): e21704.

[19] SchwingshackL L, Knüppel S, Schwedhelm C, et al. Perspective: NutriGrade: A Scoring System to Assess and Judge the Meta-Evidence of Randomized Controlled Trials and Cohort Studies in Nutrition Research [J]. Advances in nutrition (Bethesda, Md), 2016, 7(6): 994-1004.

[20] Movsisyan A, Dennis J, Rehfuess E, et al. Rating the quality of a body of evidence on the effectiveness of health and social interventions: A systematic review and mapping of evidence domains [J]. Res Synth Methods, 2018, 9(2): 224-242.

[收稿日期:2023-06-28　修回日期:2023-09-05]
(编辑　薛云)